東京電機大学

博　士　論　文

認知的満足化による強化学習アルゴリズム

Reinforcement Learning Algorithm with Cognitive Satisficing

2016 年 3 月 17 日

甲野 佑

目 次

図 目 次

表 目 次

第1章　はじめに

人工知能の中でも，ある種の主体性を持って成長していく知的エージェントには，予め決められたルールや法則からなる単純な振る舞いだけでなく，観測によって獲得した情報 (感覚刺激，報酬信号，教師信号) から，帰納的に法則を見いだして学習していく能力が必要になる．一般的には，まず選択肢の提示 (行動選択肢，モデルのバリエーションそのもの，モデルに対するパラメータ空間の定義等) があり，そしてその後に選択肢をどう選択するか，どう評価するかの問題がある．情報が未知の場合，エージェントが主体的に情報を収集し，統合し，補完して選択する事になる．環境が定常的である場合，エージェントは最適解を見出し，その最適な選択肢の組み合わせを，各 step や各状況毎の選択に収束させる事が学習の目的であると考えられる．しかしながら，環境が定常か非定常かもわからないときは，そのような選択の収束が正しいとは限らない．ここでいう環境の定常さ，非定常さについては後述するが，ある見方では観測の不完全性を前提とした際の学習についての議論には必然的に想定せざるを得ない問題環境だと言える．少なくとも学習の初期から中途において環境を正しく完全に認識し，分類できるという前提に立つ事は出来ない．本研究は，そのような定常さに関する不確実性や，完全に観測・分類する事を前提とする場合に，定常的な前提を持っている場合と比べて途端に複雑性の高まる問題環境への対応を目指す際の，非常に初歩的な議論を含んでいる．個々の選択 (あるいは意思決定という言い方が適切かもしれない) の方針には，一般的に現状で正しいと考えられるものを選ぶ利益追求的な方針と，探索的な方針があるとされる．しかし探索にも単純に未知である環境の情報を収集する目的の選択と，現環境が以前と変わっていないか確かめるための環境全体を見通す観測気球的な探索があると考えられる．一般に高度な環境適応能力を持つ人間は．既知である環境でも異常 (突然の変化等) がないかを確かめるため，その動向を図るために全体を見通すような行動をよく取る．しかしほとんどのアルゴリズムは環境が定常である事を前提としているし，そもそも非定常を前提とするための議論が整っていないため，このような定常・非定常に対する不確実さを持つ環境への対処について言及される機会は少なかった．本研究では上述の人間の柔軟さを機械学習における選択 (意思決定) に取り入れる事を目的としている．具体的には人間の因果関係の強さの感じ方と高い相関を持つ価値関数を扱い，定常か非定常かすらわからない未知なる環境における選択アルゴリズムについての提案と，そのアルゴリズムから，定常性における不確実性への対処法ついての所見を逆説的に論じる．

このような定常・非定常の不確実さは特に強化学習において顕著になる．強化学習は試行錯誤から環境から観測できる状態信号と報酬信号から，試行錯誤的に報酬を最大化する方策 (行動の系列) を獲得する事を目的としている．既存の強化学習手法は環境に対する探索はいずれ終わり，いつか最適な報酬が得られる前提をがある場合のみ，正しい行動系列が得られる事が保証されている．しかし現実的な環境ではマルチエージェントであったり，観測の不完全性から非定常である．ゆえに特に強化学習において上述のような人間の柔軟さを取り入れる事は不可欠だと考えられる．以上から，本研究ではまず定常・非定常の不確実さを持つ課題環境 (K 本腕バンディット問題) における抽象的な意思決定について論じた後，より現実的な課題での応用例として複雑なダイナミクスを持つ強化学習課題に対する提案アルゴリズムの実装とシミュレー

ションを行った.

1.1　速さと正確さのトレードオフ

エージェントが主体的に行動して環境との相互作用から学習する強化学習の枠組みでは，エージェントは利益を追求するため，むしろその時点での利益追求行動とは異なる探索行動をとる必要があるとされる．探索行動は暫定的な利益より，最終的に獲得する利益を重視し，正しい推論を行うため，なるべく多種多様な観測情報を獲得するために行われる．しかし現時点の少ない情報から利益を追求する選択ばかりしていては，主体的に新たな情報を探索する事ができない．そのためエージェントが不足する情報を探索するためには，利益を追求するための行動と適宜切り替えて意思決定される場合が多い．このような選択の仕方は，数ある行動選択肢の中で “どの行動を選ぶか” ではなく，“何のために行動を選ぶか (利益追求か探索か)” を先に決めるという選択のための選択を意味する．しかし単位時間に可能な選択肢の試行回数には物理的に限度があるため，限りある選択資源を探索に割くか，利益追求に割くかという配分に関して必ずジレンマに陥る．探索行動を多くすれば報酬獲得は後回しになり，即時的な利益追求行動を多くすれば，より大きな報酬の獲得を見逃す事になりかねない．即ち報酬獲得において，その速さと正確さにはトレードオフの関係があり両者を最大化する事はできない (Speed-Accuracy Trade-off)[Wickelgren 77]．情報探索時の選択資源を支えるリソースをどこから捻出するか問題はあるものの，静的な環境であれば最終的な正しさのみを優先し，ギリギリまで探索に選択資源を割り当てる事も間違えではない．しかし探索の優先が通用するのは環境の定常が保証される場合のみであり，動的環境ではその探索が水泡に帰す可能性は多分に存在する．エージェントが完全に観測・制御する事のできない環境はむしろ動的であると想定するべきであり，また，行動には必ずコストが伴うため，無報酬で探索し続けるにも限度がある．上述のような企業や政府等の巨大な集団で選択資源が豊富な場合の意思決定ではある程度無視できる問題も，個であるエージェントにおける意思決定は非常に繊細な問題になり得る．

しかしながら現存する実在の知能である人間は，現実の動的な環境に存在しながらも上述した利益追求と探索の切り替えを柔軟に行い，トレードオフに対処していると考えられる．この事から我々は人間の認知的特性に触発され考案された価値関数を用いる事で，エージェントの選択にトレードオフへの対処を柔軟に行う性質を付加できると考えた．以上の仮説を元に，本研究では，篠原が開発した人間の認知的な性質である対称性推論に由来する限定的な主観確率モデル，Loosely Symmetric model (以下 LS)[篠原 07] に関する考察から，我々が新たに開発した認知的なアルゴリズム，EXtended Loosely Symmetric model (以下 LSX)[甲野 14] について，トレードオフへの対応という観点から各種性質を示す.

1.2　環境と選択収束

エージェントの理想的な学習は，隠れ変数を含めた環境の状態を正しく認識でき，全ての状態に対して最適な行動を取れるという，状態と選択の対が全て最適である学習結果を目指すものである．しかし，複雑な環境で生きるエージェントにとって，学習とは終わるものではない．環境の全ての状態は (近似的にも) 把握しきれるものではないし，各々への最適な選択はさらに難しい要求となる．そのため特定のモデルを用いて環境を近似し，その選択を司るパラメータを最適化する方策が取られるわけだが，これもモデルを環境として，パラメータの決定を選択とするならばほぼ同じ問題が適用できる．むしろ何かしらの近似を行えば非定常性への対処が

難しくなる場合がある (環境のわずかな差異がつぶされるため). ゆえに機械学習では過渡的には環境の非定常性を考慮して学習を行うべきである.

定常環境での学習では, エージェントが有する解 (の選択) が十分収束したと言える場合や, あらかじめ規定された時間の学習フェイズが終了した時に学習が終わったと言える. しかし非定常環境では, 時と場合によって解を変化させ続ける事が重要であるため, 学習が終了したという状況を規定しにくい. これが前述の議論で選択という言葉を多用していた理由である. 非定常環境では程度の差こそあれ, 学習し続けなければならず, また同時に成果を出し続ける事が重要であるため, これらを総合して “選択” という単語を用いた. そして本論文において, 我々は議論の簡略化のため, 定常環境で学習の終了時に見られるエージェントが特定の状態-選択肢対のパターン (強化学習における方策に相当) を採用し続けるような選択の強い偏りを “選択収束” と呼ぶ. 本論文ではシンプルな強化学習課題についてしか扱わないが, 学習器等のパラメータの選択についても同様の見方を当てはめる事ができると考えらえる. つまり以上のような定常か非定常かも不明な環境における選択課題において, 局所的な収束を防ぎ最適な選択収束へ向かいつつ, 環境の変化を考慮しつづける事が理想的なエージェントの選択方策であると言える.

第2章　Loosely Symmetric model

人間は「一を聞いて十を知る」という言葉の通り，獲得した情報を加工して，獲得した情報以上の知識を見いだす事ができる．この様な情報の “汎化” は，事前に獲得した知識と組み合わせた演繹的推論や，人間が持ち合わせる非論理的推論傾向の結果であると考えられる [Hattori 07]. 社会生活における人間に必要な情報や知識は膨大であるのだが，それに対して得られる情報の種類は稀少であると考えられる．特に未知の環境 (転居先，新任の職務) においては，その環境から得られる情報は非常に稀であり，しかしながら，その構成員を含めた環境からは，なるべく速い適応を要求される事になる．このように人間は必要な知識に対して常に不確実で粗い情報しか得る事ができず，それゆえに不確実さに対応するための方法を有していると考えられる．

篠原 [篠原 07] は人間の非論理的な推論形式である緩く対称性推論を行う性質を確率に拡張した主観的な信念形成モデルとして Loosely Symmetric model(以下 LS，式 2.1) を考案した．このモデルは 2 事象 p，q 間に対する人間の直感的な因果関係の強さの推定 (因果帰納) と高い相関を持つ．また未知の選択肢に対して，試行/未試行とそれに対する報酬の観測から，獲得報酬量を最大化する事を目的とする単純な意思決定課題，2 本腕バンディット問題において高い成績を有する事が示されている [篠原 07].

$$LS(q|p) = \frac{P(p \cap q) + B(\bar{q}; p)}{P(p) + B(q; p) + B(\bar{q}; p)} \tag{2.1}$$

$$B(\bar{q}; p) = P(\bar{q})P(\bar{p}|\bar{q})P(p|\bar{q}) \tag{2.2}$$

$$B(q; p) = P(q)P(\bar{p}|q)P(p|q) \tag{2.3}$$

因果推論と意思決定課題の双方に対して同一の評価モデルが高い成績を有するモデルは他に非常に稀である [Takahashi 10]．我々はこのモデルに対して分析を通して，少ないサンプル数の下，即ち上述の情報に対する飢餓状態の際に，抑制された評価 (情報を安易に信じない，他と比較する) をする点が意思決定課題や人間の認知特性と類似するのでは無いかと考えた．LS は客観的な条件付き確率 $P(p|q)$ に認知的なバイアスとして定義される変数 (式 2.2，式 2.3，以後バイアス項) を加える事で定義される信念形成のモデルであり，要約すると「事象 p によって事象 q が生じる」という命題を信じる主観的な強さを意味する．また，重み変数として式 4.7 を定義すると，LS を $P(q|p)$ と $V_B(\bar{q}; p)$ の重み付き平均と見なす事も出来る．

$$\omega_B = \frac{P(p)}{P(p) + B(\bar{q}; p) + B(q; p)} \tag{2.4}$$

$$V_B(\bar{q}; p) = \frac{B(\bar{q}; p)}{B(\bar{q}; p) + B(q; p)} \tag{2.5}$$

$$LS(q|p) = \omega_S P(q|p) + (1 - \omega_S)V_B(\bar{q}; p) \tag{2.6}$$

高橋 [Takahashi 10] は前件事象 p，$\bar{p}$ に対して LS のバイアス項が不変 ($B(q; p) = B(q; \bar{p})$ かつ $B(\bar{q}; p) = B(\bar{q}; \bar{p})$) である事を，視覚にける非着目対象 (地) の印象が着目対象 (図) に依存しない事に準えて，“地の不変性” と呼んだ．前述の対称性推論の性質を有する信念形成モデルと

いう条件だけでは，生成可能なモデルは非常に膨大な数になる．その中で *LS* が対称性推論のモデルとして妥当だと見なせるのは，地の不変性という認知に由来する性質を有する事と因果的直感と高い相関を持つ点からである [Takahashi 10]．*LS* が何故，上述のような良い振る舞いをするのかは対称性推論と関係のあるとされる，人間が意思決定の際に用いるとされる “競合”[Tversky 74]，“信頼性考慮”[Kahneman 84]，“満足化”[Simon 56] という三つの性質との関連に由来するとされる．

2.1　競合

対称性推論では「p ならば q である $(p \to q)$」から「p でないなら q でない $(\bar{p} \to \bar{q})$」を導く．この事を意思決定課題に当てはめると,「ある選択肢 a から結果 e が生じた $(a \to e)$」から「その他の選択肢 $\bar{a}$ からは結果 e が生じない $(\bar{a} \to \bar{e})$」を導く事になる．これを確率に拡張すると，選択肢 a を行い結果 e が生じる確率 $P(e|a)$ が高くなるに従い，連動して選択肢 a 以外から選択肢 e が生じる確率 $P(e|\bar{a})$ が低くなる事を意味する．これは他により良い選択肢がある場合，他の価値を相対的に低く見積もる事によってその発見を疎外する．しかし同じ仕組みで「ある選択肢 a から結果 e が生じなかった $(a \to \bar{e})$」から「その他の選択肢 $\bar{a}$ からは結果 e が生じない $(\bar{a} \to e)$」を導く事もできる．こちらは $P(e|a)$ が低くなると，それに連動して $P(e|\bar{a})$ が高くなる事を意味するため，その他の選択肢に対する評価を上げる事で，他の可能性に目を向ける事が可能になる．即ち，その他の価値を上げる事によって探索的行動を促す効果がある．

このように相対的に評価する事は，対称性推論から $a \to e$ という情報を $\bar{a} \to \bar{e}$ として最大限に扱う事で，少ない試行回数を補う効果を持つと考えられる．また一方の評価と反対の評価を与える事で，探索を行うための評価の逆転を促す事が可能となる．また，結果 e をもたらすのが a と $\bar{a}$ の間で二者択一の構造を持ち，それぞれで価値を奪い合う事からこの性質を “競合” と呼ぶ事にする．このような競合において価値というのは「その選択肢がどの程度，結果に結びつくか」のみだけでなく「その選択肢にどの程度注視するか」という意味合いが含まれることになる．任意の選択肢 a が結果 e を発生し続ければ，即時的な利益の追求行動が促され，逆に結果が生じなければ他の選択肢に注視を向ける探索行動が促される．この性質により，速さと正確さのトレードオフを持つ課題に対して，実際の試行結果から速さを重視するか，正確さを重視するかを自律的に調整しているのだと考えられる．即ちこれは人課題の提供する環境に併せて，その解法となる方策を柔軟に変化させる性質の源が競合にあり，更にそれは獲得した情報を対称性推論によって価値と注目の二重の意味として考慮する事により記述可能である事を示している．

2.2　信頼性考慮

信頼性考慮は上述の情報に対する不足の度合いを定量化して評価に反映させる事である．即ち情報サンプルが不足する情報に対する評価を抑制的に行う性質である．選択肢 a を試行して得られた $a \to e$ という情報と，選択肢 a 以外の試行結果 $(\bar{a} \to \bar{e})$ に対する対称性推論の結果で得られた $a \to e$ という情報を，価値に差をつけず扱うならば，両者を区別する事ができない．つまりこの場合は試行回数が多い選択肢の結果が他の結果に対する影響力を強く持ってしまう．ある選択肢 a への情報が他に比べて充分にあれば選択肢 a に対する評価は正確になる．しかし不足していれば他の選択肢からの影響によって，選択肢 a のみに対する観測結果がないがしろ

にされる．対称性推論はこのように対象の観測頻度を評価に反映させるという点で信頼性考慮していると考えられる．

LS における信頼性考慮の形式は，対称性を緩める性質を持つためもう少し複雑である．対称性推論から $P(e|a)$ に対して $P(\bar{e}|\bar{a})$ を同様の意味として扱う際にその価値に差がないと考えれば，両者の影響度の比率は観測全体に対する観測割合である $P(a) : P(\bar{a})$ となるはずである．しかし *LS* は $P(a) : P(a|e)P(\bar{a})$ という比率になり，対称性推論によって生み出した情報を試行回数に対して $P(a|e)$ 分だけ減衰して重み付けする．この減衰により *LS* は緩い対称性推論を実現している．

2.3 満足化

満足化とは目的志向型の意思決定課題における方策を与える物で，ある具体的な基準を達成するよう試行錯誤する事を指す [Simon 56]．つまり満足化という方策の下では試行錯誤の目的が「最も良い価値を持つ選択肢を見つける」ではなく,「(ある具体的な) 基準を満たす選択肢を見つける」になっている．基準値を行動コストと釣り合っていて充分に満足するための指標と考えれば，意思決定課題において妥当な方策である定義の難しい「最高の選択肢」を探す労力と比較すれば，速さと正確さのトレードオフへの対応という観点でも非常に有用なことがわかる．

しかしながら満足化のみで対称性推論との関係を述べるのは難しく，上述の競合と信頼性考慮の議論を利用する．まず競合では一方の選択肢の評価がそれ以外の選択肢に対する逆の評価を引き起こし易くすると述べた．ここで選択肢が a_1 と a_2 の二つしか無い場合 ($A = \{a_1, a_2\}$, $\bar{a_1} = a_2$)，選択肢 a_1 に関する評価が下がるような試行結果 ($a_1 \rightarrow \bar{e}$) を観測した時，競合の上では結果から導かれる情報 ($a_2 \rightarrow e$) によって同時に選択肢 a_2 が同じ分だけ評価を上げる．そのまま選択肢 a_1 への評価が下がり続けると，同じ速度で a_2 への評価が上がり続け，どこかで評価が逆転する．ここで定義する相対評価は，全ての選択肢に対する初期の評価値の総量を取り合う形になるため，選択肢が二つの場合は，この評価が逆転する値は任意の値を取る．即ち，その値が上述で定義した基準値となる．それはある選択肢に対する評価値が評価が逆転する値を下回った時，必ずもう一方の選択肢がこれを上回るためである．

LS ではもう少しこの性質がわかり易く表れる．*LS* では観測割合 $P(a_i)$ の極限を取る事により，式 4.12 や式 4.13 になるという性質がある．即ち，ここでは 0.5 が基準値となっており，最も *LS* の値が高い選択肢を選ぶようなアルゴリズムでは「現在の観測から結果 e を生じる確率が 0.5 より高い選択肢 a_i を探す」目的が与えられることになる．

$$\lim_{P(a_1)\rightarrow 1.0} LS(e|a_1) \approx P(e|a_1) \tag{2.7}$$

$$\lim_{P(a_1)\rightarrow 0.0} LS(e|a_1) \approx 0.5 \tag{2.8}$$

LS はこれらを信頼性考慮によって柔軟に切り替える事で，確率的な価値の推定モデルとして実装されており，これが因果帰納と意思決定の両分野でうまく働いている理由であると考えられる．

第3章　K本腕バンディット問題

本研究では前述した選択，あるいは意思決定に関する初歩的な議論を行うために，K本腕バンディット問題を例に，何も情報の無い状態から, トレードオフを抱える課題，環境に対し主体的に情報を獲得して行く際の不確実な知識の扱い方や評価方法について論じる. ここでの不確実な知識とは観測が不十分で, 正しいか否か断定出来ない曖昧な知識を意味する. これは強化学習課題における初期において学習を促進するためにどのような方策や価値観数を用いるかの問題に対応する [Sutton 00]. K本腕バンディット問題とは目的となる報酬を確率的に得る事の出来る幾つかの手段集合 $A = \{a_1, a_2, ..., a_n\}$ から最適な手段を探索し, 得られる報酬 $E = \{e, \bar{e}\}$ を最大化させる事を目的とする問題である. K本腕バンディット問題は意思決定課題の一種であり, 動物が餌獲得のため複数の餌場から餌獲得が最も期待される餌場を探索し, 決定する課題に例えられる. この課題の難しさは探索と収穫のジレンマという単語で表される. 目的を達成するための手段の中から最も効率の良い手段を知るためには, 情報探索のための試行に多くの時間を費やさなければならない. それは結果的に見れば, 非効率的な手段を何度も行った事になり, 探索のための試行を行えば行う程, 最終的に得られる報酬は低くなる. しかし, 探索が不十分だと正確な確率を知る事が出来ず, 不幸にも偶然それまで報酬が多く得られていただけの非効率的な手段を効率的であると判断を誤ってしまう可能性が高くなる. 生き物が効率的に生きるためには, 度々このようなバンディット問題的な課題に直面する. 例えばある野良猫にとって数カ所の餌場があったとする. それらの餌場を訪れると確率的に餌を得られるが, 時間的制約があるため全ての餌場を廻る事は出来ない. 餌場を腕, 餌を報酬とした時に, これは正にバンディット問題に置き換える事が出来る. 現実において, バンディット問題における対応すべき環境は複雑であり非定常である. 猫にとって餌場の価値は, 餌を出していた住人が突然病で餌を出せなくなったり, 移住で猫好きの住人に変わり, 餌を出す頻度が増したりする等で突然変化する事が有る. あまつさえギャンブルのスロットマシンでさえ, 時間や試行手順によって報酬獲得の確率が変動する場合がある. そのような非定常な環境に対し, 複雑な準備をせず, かつ早く簡便に対応するのは難しい. この課題の難しさは前述した探索と利益追求のジレンマという単語で表す事が出来る．高い報酬を得るためにはどこかで探索を辞めるべきである. しかし探索しなければ高い報酬を得る事はできない．K本腕バンディット問題はこのような知識の獲得とその利用からなる普遍的な“速さ”と“正確さ”のトレードオフを端的に表す事が出来る課題である.

3.1　観測情報の表現形式

K本腕バンディット問題において1回の試行毎に得られる情報は，どの選択肢を試行したか ($a_{Select} = a_i$) と報酬が得られたか否か ($e_{Gain} = e$ or $\bar{e}$) である．そのため，K本腕バンディット問題で得られた情報の蓄積に対する割合的な表現は 3.1 に表される形式になる．これは完全結合分布の形式を取っており，周辺化を行う事で任意の選択肢を試行した割合 $S(a_i)$　や，報酬を得られた割合 $S(e)$ が計算できる．更にその選択肢を試行して報酬が得られた割合 (サンプル平均) である条件付き割合 $S(e|a_i)$ も計算でき，多くのアルゴリズムはこの条件付き割合 $S(e|a_i)$

表 3.1: 事象 A，E 間の完全結合分布

	$e(win)$	$\bar{e}(lose)$
a_1	$S(a_1 \cap E)$	$S(a_1 \cap \bar{E})$
a_2	$S(a_2 \cap E)$	$S(a_2 \cap \bar{E})$
$\vdots$	$\vdots$	$\vdots$
a_n	$S(a_n \cap E)$	$S(a_n \cap \bar{E})$

表 3.2: 事象 A，E 間の頻度分布

	$e(win)$	$\bar{e}(lose)$
a_1	w_1	l_1
a_2	w_2	l_2
$\vdots$	$\vdots$	$\vdots$
a_n	w_n	l_n

をその選択肢 a_i の報酬に対する価値として扱い，選択肢を選ぶための情報として参照する．

3.1 は確率的な表記であるため，一回の試行毎に全ての $S(a_i \cap e)$，$S(a_i \cap \bar{e})$ を更新する事になる．しかしそれは計算上では非効率であるし，アルゴリズムの概要を把握し難いため，実際の計算では 3.2 に示す頻度表記を用いる．

頻度表記である 3.2 は，確率表記である 3.1 の各セルに対して総試行回数 n_{sum} を積算することで導出できる．変数 w_i は任意の選択肢を試行 (a_i) して 報酬を得た (e) 場面をどのくらい観測したかを意味する量的な表現で，変数 l_i は逆に任意の選択肢を試行 (a_i) して 報酬を得られなかった $(\bar{e})$ 場面に対応する．後述する更新方法によっては変数 w_i，l_i は純粋な試行回数と呼べない場合があるので，ここでは観測量と呼ぶ．また式 3.1 や式 3.2 等により，任意の選択肢を試行した割合 $S(a_i)$ や条件付き確率 $S(e|a_i)$ 等の確率表記による指標も同様に計算できる．

$$S(a_i) = \frac{w_i + l_i}{\sum_k (w_k + l_k)} \tag{3.1}$$

$$S(e|a_i) = \frac{w_i}{w_i + l_i} \tag{3.2}$$

頻度表記である 3.2 では，試行毎に試行した選択肢 a_i とその結果 e or $\bar{e}$ に該当する任意のセル情報を更新するのみで更新が完了するため，非常に節約的であり議論もシンプルになる．しかし以下では必要に応じて確率表記も使い分ける．

3.2　収穫と探索のジレンマ

前述の通り，意思決定課題である K 本腕バンディット問題は報酬の最大化とそのための情報探索の間にどちらを優先すべきかのジレンマが生まれる．単純な見解として報酬の最大化とは，現在得られている知識において最も価値が高いと推定される手段を選択し続ける事を意味し，greedy な行動に対応する．対して探索とは現在得られていない知識を獲得するために，あえて非 greedy な行動を取る事で実現される．K 本腕バンディット問題に用いられる代表的なア

ルゴリズムとして ϵ - greedy 法が存在する．これは確率 ϵ で非 greedy なランダム選択を行い，確率 $1-\epsilon$ で客観的報酬確率に基づいて greedy に報酬の最大化を行う．つまり報酬の最大化と情報探索に対応して，それらを乱数によって完全に分離する最も簡潔なアルゴリズムである．しかしながら，報酬の最大化と情報探索は完全に分離するべきかどうかには議論の余地がある．たしかに収穫と探索の間にはジレンマがあるかもしれないが，意思決定課題を解くにあたっては，なるべく報酬が多く得られるように探索したり (例えば探索する際でも高い価値を持つ選択肢を多めに探索する等)，明らかに低い選択肢を考慮から外したり，ある程度は利益追求を考慮しつつ効率的に探索する事が重要であるように思われる．この様な効率的な探索と言うアプローチを有する選択方策として， Boltzmann 分布を用いた softmax 法が存在する．具体的には，選択肢それぞれの価値 $P(e|a_i)$ の高さに応じて，価値が高いほどに選択肢が選ばれる確率が上昇するという選択方策である．しかし softmax も ϵ - greedy 法と同様に，乱数を用いてある割合で報酬の最大化と探索のどちらを行うか決定している点では同様である [Sutton 00].

ϵ-greedy 法，softmax 法では観測によって得られた客観的な報酬割合 $S(e|a_i)$ を評価に用いているが，現在得ている情報に基づき価値を推定する場合は，それ以外の方法を考える事も出来る．特に人間が行う価値の推定には必ずと言って良い程主観的な偏りを生じる．その様な，偏りを持つ価値の推定方法を扱う場合，価値に対する greedy な行動は，必ずしも常に報酬を求めた貪欲な行動であるとは言えず，そこには探索的意義も含むようになる．LS による評価はそのような意味合いを持つと考えられる．対して，K 本腕バンディット問題において高い成績を示す UCB1 アルゴリズムは統計的な指標によって，価値を推定している [Auer 02]．UCB1 は以下の式 3.3 により任意の手段 a_i に対する価値を表現し，その価値に対して常に greedy に行動する．

$$S(e|a_i) + \sqrt{\frac{2\ln n_{sum}}{n_i}} \tag{3.3}$$

第一項は客観的に得られた手段 a_i の報酬確率であり，第二項に出現する変数 n_i は手段 a_i の試行回数，変数 n_{sum} は全体の試行回数を意味する．UCB1 は第二項によって試行回数からその手段に関する相対的な知識量を評価に含め，知識の少ない程に価値を高く推定する．即ち統計的な指標に基づき，観測された結果から相対的な区間推定を行い，その上限を価値としていると言える．より抽象的な表現をすれば UCB1 アルゴリズムは試行結果を常に「運が悪かった」と評価して，価値の底上げを行っている．そのため，報酬の最大化と探索の度合いは第 1 項，第 2 項の間のバランシングで決定し，ϵ-greedy 法や softmax のように乱数を用いず，試行結果を楽観的に再評価する事により， greedy に行動するのみで試行に双方の意味合いを持つ事が出来る [Auer 02]．UCB1 アルゴリズムは機体損失の上限が保証されているという優れた性質を持つ．言い換えれば，長期的には必ず最も高い報酬確率を持つ選択肢を見つけ出すという性質である．これは速さと正確さのトレードオフにおいて，探索と利益追求を明らかに異なる行動として純粋に分離する事が困難，あるいは分離しない方が良い成績を示すという示唆を与えている．しかし機体損失の上限が保証されているという事は， UCB1 アルゴリズムが統計的な指標を用いる事によって，常に効率的な情報探索を行っているアルゴリズムとも言える．つまり長期的な利益につながる正確さを犠牲にしてでも，問題に応じて速めに報酬を獲得したいという，速さと正確さのトレードオフに対する根源的な問題には対応していないという事になる．

3.3　バンディット問題における選択収束

本論文では議論を簡略化するために，いずれかの選択肢の選択された割合がほぼ 100 %になる状態を “選択収束” と定義する ($\max(S(a_i)) \approx 1.0$)．言い換えると，その時点である選択肢に執着して他の選択肢を相対的に殆ど選択していない状態を意味する．その執着している選択肢が真に最も期待値の高い正解の選択肢である場合，後悔の度合いの上昇が止まり，上限が決定する．逆に期待値が最も高くない誤った選択肢に執着してしまっている場合，その状態から抜け出せなければ後悔の度合いは上昇し続けてしまう．

第4章　EXtended Loosely Symmetric model

上述した価値関数 LS は 2 本腕バンディット問題において，LS を各選択肢に対する価値関数とし，“最も高い LS 値を持つ選択肢を選択していくのみ” で探索行動を意識せずとも高い成績を得られる事が示されている [篠原 07]．これは UCB1 と同様に，選択の履歴に基づいて各々の選択の重要性を評価し，価値関数の空間が主体的に変化していくためだと考えられる．K 本腕バンディット問題における LS は選択肢 (a_i) と結果 (e) の間の結びつきの強さとして以下の式で定義される．

$$LS(e|a_i) = \frac{w_i + b(\bar{e};e_i)}{w_i + l_i + b(e;e_i) + b(\bar{e};e_i)} \tag{4.1}$$

$$b(\bar{e};a_i) = \frac{l_i(\sum_k l_k - l_i)}{\sum_k l_k} \tag{4.2}$$

$$b(e;a_i) = \frac{w_i(\sum_k w_k - w_i)}{\sum_k w_k} \tag{4.3}$$

あるいは観測割合を意味する表現 S を用いて以下のようにも表せる．

$$LS(e|a_i) = \frac{S(a_i \cap e) + B(\bar{e};e_i)}{S(a_i) + B(e;e_i) + B(\bar{e};e_i)} \tag{4.4}$$

$$B(\bar{e};a_i) = \frac{S(a_i \cap \bar{e})S(\bar{a_i} \cap \bar{e})}{S(\bar{e})} \tag{4.5}$$

$$B(e;a_i) = \frac{S(a_i \cap e)S(\bar{a_i} \cap e)}{S(e)} \tag{4.6}$$

因果推論と意思決定課題の双方に対して同一の評価モデルが高い成績を有するモデルは他に非常に稀である [Takahashi 10]．LS は客観的な条件付き割合 $S(p|q)$ に認知的なバイアスとして定義される関数 (割合的表現では B，頻度的表現では b，以後バイアス項) を加える事で定義される信念形成のモデルであり，要約すると「選択 a_i によって結果 e が生じる」という命題を信じる主観的な強さを意味する．また，重み変数として式 4.7 を定義すると，LS を $S(e|a_i)$ と $V_B(\bar{e};a_i) = 1 - V_B(e;a_i)$ の重み付き平均と見なす事も出来る．

$$\omega_B = \frac{S(a_i)}{S(a_i) + B(\bar{e};a_i) + B(e;a_i)} \tag{4.7}$$

$$V_B(\bar{e};a_i) = \frac{B(\bar{e};a_i)}{B(\bar{e};a_i) + B(e;a_i)} \tag{4.8}$$

$$V_B(e;a_i) = \frac{B(e;a_i)}{B(\bar{e};a_i) + B(e;a_i)} \tag{4.9}$$

$$1 = V_B(e;a_i) + V_B(\bar{e};a_i) \tag{4.10}$$

$$\begin{aligned} LS(e|a_i) &= \omega_B P(e|a_i) + (1-\omega_B)V_B(\bar{e};a_i) \\ &= \omega_B S(e|a_i) + (1-\omega_B)(1-V_B(e;a_i)) \end{aligned} \tag{4.11}$$

高橋 [Takahashi 10] は前件事象 p，$\bar{p}$ に対して LS のバイアス項が不変 ($B(q;p) = B(q;\bar{p})$ かつ $B(\bar{q};p) = B(\bar{q};\bar{p})$) である事を，視覚にける非着目対象 (地) の印象が着目対象 (図) に依存しない事に準えて，“地の不変性” と呼んでいる．また，LS は大用 [大用 15] により，満足化と呼ばれる意思決定における人間の選択傾向を有する価値関数であると言われている．

4.1　K 本腕バンディット問題における LS の満足化

K 本腕バンディット問題において LS を価値関数として選択に用いると，ただ貪欲に選択 (Greedy，価値関数が最も高い選択肢を選ぶ事を意味する) するのみで満足化と同様の傾向が得られる．具体的には LS の a_i に対する観測割合 $S(a_i)$ の極限を取る事により，式 4.12 や式 4.13 に収束する性質から導かれる．

$$\lim_{S(a_i)\to 1.0} LS(e|a_i) \approx S(e|a_i) \tag{4.12}$$

$$\lim_{S(a_i)\to 0.0} LS(e|a_i) \approx 0.5 \tag{4.13}$$

極限として $S(a_i) \approx 1.0$ は行動 a_i に対して選択収束が起こっている様を示し，相対的にある行動 a_i に対する知識が豊富で，a_i に強く注目して来た事を意味する．その逆に $S(a_i) \approx 0.0$ は他の行動に選択収束が起こっている様や，ある選択肢 a_i に関する知識がほとんど無い事を示し，あまり注目されてこなかった事を意味する．2 本腕バンディット問題において LS を価値関数として用いた場合，選択収束が起こると選択収束の収束対象になっている行動選択肢 a_{cnv} は式 4.12 から $S(e|a_{cnv})$ という客観的に観測された報酬獲得割合に収束する．一方，収束対象になっていない行動選択肢 $\overline{a_{cnv}}$ は式 4.13 から 0.5 という固定値に収束する．我々は前者のような選択収束された事に起因する評価値の客観化を “鮮明化”，後者の様な情報の相対的な不足に起因する評価値の固定化を “背景化” と呼ぶ事にする．選択収束していくと，ある選択肢の評価が先鋭化し，それに連動してもう一方の背景化が起こる．しかし，先鋭化された価値 $S(e|a_{cnv})$ が 0.5 より低い場合 ($S(e|a_{cnv}) < 0.5$)，もう一方の選択肢 $\overline{a_{cnv}}$ が選択されてしまう．最初から LS を用いて選択をしていく場合，選択収束の際には $S(e|a_{cnv}) > S(e|\overline{a_{cnv}})$ であるため，これは客観的価値に照らして探索と定義できる．行動 a_{cnv} に対して選択収束が起こっている場合に $\overline{a_{cnv}}$ が選ばれてしまうと，選択収束に逆らうため，鮮明化と背景化が抑制される．逆に鮮明化された価値 $S(e|a_{cnv})$ が 0.5 より高い場合 ($S(e|a_{cnv}) > 0.5$)，そのまま行動 a_{cnv} が選択され続け，選択収束がより進んで行く．これを端的に表現すると，LS を価値関数に用いた選択では 0.5 が基準値となっており，最も LS 値が高い選択肢を選ぶように設定するだけで「現在の観測から結果 e を生じる確率が 0.5 より高い選択肢 a_i を探す」目的が与えられると言える．この様にある基準を満たす選択肢を発見するまで探索を止めず，見つけたら探索を止める事から，LS を用いた選択は基準 0.5 に対する満足化を含んでいると言える．

4.2　LS の問題点とその改善

価値関数としての LS には，そのままではより一般的な意思決定課題に対応できないという課題があった．おおまかには以下の二点で通りである．

1. 3 つ以上の選択肢に対応できない
2. 参照点を任意に設定できない

問題点 (1) は，*LS* がそもそも二要因間の生起・不生起のみに着目した信念形成モデルである事に起因する．即ち，K 本腕バンディット問題においてそれを解り易く表現するならば,「行動 a_i によって結果 e が生じる」事への結びつきの強さを意味する．この定義に従えば，通常，選択肢の数が幾つであろうとも「任意の行動 a_i によって結果 e が生じる」事の結びつきの強さは計算可能である．しかしながら，従来の研究からその単純なアプローチでは 3 つ以上の選択肢を持つ K 本腕バンディット問題では良い成績を得る事が出来なかった．即ち，これは K 本腕バンディット問題においての *LS* の定義にそもそもの問題があり，3 つ以上の選択肢が扱えないのだと推察される．

問題点 (2) は，*LS* において定義される基準値が 0.5 に固定されており，それを任意に変更する事が出来ない事を意味する．K 本腕バンディット問題における選択肢群の真の報酬獲得確率からなる問題環境と基準値によって *LS* が得意な環境，不得意な環境が定義される．つまり基準値を変更するメリットとは，問題環境に合わせて *LS* にとって得意な環境になるようチューニングする事が可能になる点にある．更に動的な試行錯誤からの観測によって問題環境を仮に定義し，それと対応する参照点を学習する事が可能になれば，*LS* を常に得意な問題環境において用いる事が可能になると考えられる．以下では上記の問題点に対する改善案を新たに得られた知見から提案していく．

4.2.1 問題点 1：3 つ以上の選択肢に対応できない

従来の *LS* は選択の在・不在，即ち 2 通りの選択肢に対してのみしか扱えなかった ($A = \{a, \bar{a}\}$)．前述した通り，元々 *LS* は「事象 p によって事象 q が生じる」事に対する強さを意味するモデルである事に起因している．K 本腕バンディット問題においても「任意の行動 a_i によって結果 e が生じる」事に対する強さとして計算する事は式 2.1 において可能であるが，その成績は非常に低かった．そのため特定の選択肢の在・不在という表現ではない，より一般化された複数の選択肢に対する *LS* の定義が模索されていた．我々はこのような *LS* の再定義に際し，*LS* の持つ特有の能力の中から，地の不変性と呼ばれる性質に着目した．前述の通り，地の不変性とは前件事象 (K 本腕バンディット問題では行動) a_i，$\bar{a_i}$ に対して *LS* のバイアス項が不変 ($b(e; a_i) = b(e; \bar{a_i})$ かつ $b(\bar{e}; a_i) = b(\bar{e}; \bar{a_i})$) である事を指す [Takahashi 10]．選択肢が 2 つ ($A = \{a_1, a_2\}$) でありそれらが排反であるならば，行動 a_1 にとっての補事象 $\bar{a_1}$ は行動 a_2 に対応するため ($\bar{a_1} = a_2$，逆もまた然り)，行動 a_1，a_2 間のバイアス項の不変は保たれている．しかしながら選択肢が 3 つ ($A = \{a_1, a_2, a_3\}$) になると，a_1 の補事象は $\bar{a_1} = a_2 \cap a_3$ となり，行動 a_1，a_2，a_3 の間にバイアス項の不変は成り立たなくなる．

我々はこの点に従来の *LS* が 3 つ以上の選択肢において上手くいかなかった原因があると考え，バイアス項の不変 (地の不変性) が保たれるような定義を提案する．

4.2.2 改善点 1：LSK

ここでは全ての任意の行動 a_i に対して不変になるようなバイアス項を考案する事を目的としている．しかしながらそれだけでは形式は決まらないため，*LS* の持つ多数の性質の中から特に重要だと考えられる性質の幾つかが保たれるようバイアス項の定義の制限に加える．LS の持つ数値的な性質は以下の二種に大別される．

1. バイアス項の不変 (地の不変性)
2. 背景化時の参照点への近似

上述したが，選択収束していくと LS はそれに伴い，収束された選択肢の評価は鮮明化 (式 4.12) されていき，そうでない選択肢は背景化 (式 4.13) されていく．LS にこのような収束が起こる理由は，LS のバイアス項 (式 4.5，式 4.6) が選択肢 a_i の観測割合 $S(a_i)$ の極限に対して以下のような性質を持つためである．

$$\lim_{S(a_i)\to 1.0} B(\bar{e}; a_i) \approx S(\bar{e} \cap a_i) \tag{4.14}$$

$$\lim_{S(a_i)\to 1.0} B(e; a_i) \approx S(e \cap a_i) \tag{4.15}$$

ここで複数の行動選択への対応へと主題を戻すと，LS の基本的な振る舞いから「最も知らない選択肢 (式 4.17) をより曖昧な評価に (式 4.19)」と「最も知っている選択肢 (式 4.16) をより客観的な評価に (式 4.18)」という性質が，複数の行動選択でも重要であると考えられる．

$$a_{MT} = \arg\max_{a_j} P(a_j) \tag{4.16}$$

$$a_{LT} = \arg\min_{a_j} P(a_j) \tag{4.17}$$

$$\lim_{S(a_{MT})\to 1.0} LSK(e|a_{MT}) \approx P(e|a_{MT}) \tag{4.18}$$

$$\lim_{S(a_{MT})\to 1.0} LSK(e|a_{LT}) \approx 0.5 \tag{4.19}$$

ここで選択肢が二つの際と同様に，式 4.19 の収束を得るためにはバイアス項の極限は式 4.20，4.21 とならなければならない．

$$\lim_{S(a_{MT})\to 1.0} B_K(\bar{e}) \approx P(\bar{e} \cap a_{LT}) \tag{4.20}$$

$$\lim_{S(a_{MT})\to 1.0} B_K(e) \approx P(e \cap a_{LT}) \tag{4.21}$$

この際に極限として定義される $S(a_{MT}) \to 1.0$ は前述した選択収束している様を意味する．選択収束する際に式 4.20，4.21 を導く事が可能なバイアス項の形式は以下が考えられる．

$$B_K(\bar{e}) = \frac{S(\bar{e} \cap a_{LT}) S(\bar{e} \cap a_{LT})}{S(\bar{e} \cap (a_{MT} \cup a_{LT}))} \tag{4.22}$$

$$B_K(e) = \frac{S(e \cap a_{LT}) S(e \cap a_{LT})}{S(e \cap (a_{MT} \cup a_{LT}))} \tag{4.23}$$

この様な極限が得られるのは選択収束に際して式 4.24，4.25 が成り立つからである．単純に選択収束時に式 4.20，4.21 を満たすだけなら，式 4.26，4.27 でも満たす事が出来るが，その収束は式 4.28，4.29 に由来する．これらは式 4.24，4.25 に比べて，選択肢が増大である場合に選択収束に対する収束の感度が悪いため，バイアス項には式 4.22，4.23 を採用した．

$$\lim_{S(a_{MT})\to 1.0} \frac{S(\bar{e} \cap a_{LT})}{S(\bar{e} \cap (a_{MT} \cup a_{LT}))} \approx 1.0 \tag{4.24}$$

$$\lim_{S(a_{MT})\to 1.0} \frac{S(e \cap a_{LT})}{S(e \cap (a_{MT} \cup a_{LT}))} \approx 1.0 \tag{4.25}$$

$$S(a_{MT}|\bar{e}) S(\bar{e} \cap a_{LT}) \tag{4.26}$$

$$S(a_{MT}|e) S(e \cap a_{LT}) \tag{4.27}$$

$$\lim_{S(a_{MT})\to 1.0} S(a_{MT}|\bar{e}) \approx 1.0 \tag{4.28}$$

$$\lim_{S(a_{MT})\to 1.0} S(a_{MT}|e) \approx 1.0 \tag{4.29}$$

これらの新たなバイアス項である式 4.22，4.23 は地の不変性を満たす．また，これらの式を頻度表現すると式 4.30，4.31 となる．

$$b_K(\bar{e}) = \frac{l_H l_L}{l_H + l_L} \tag{4.30}$$

$$b_K(e) = \frac{w_H w_L}{w_H + w_L} \tag{4.31}$$

$$LSK(E|A_i) = \frac{w_i + b_N(\bar{E})}{w_i + l_i + b_N(\bar{E}) + b_N(E)} \tag{4.32}$$

また，K 本腕バンディット問題の選択肢が 2 つの時は *LS* と *LSK*(式 4.32) は完全に一致する．

4.2.3 問題点 2：参照点を任意に変更できない

前節においても議論に上がったが，*LS* における式 4.13，*LSK* における式 4.19 のように，*LS* の基本的な振る舞いとは「知らない選択肢の評価を曖昧な基準値に近似する」という性質である．これは前述した満足化において述べた通り，よく知らない選択肢が基準値に近似される事で「基準値以上に評価できる選択肢が無い」のならば「よく知らない選択肢を試行してみよう」という探索的選択が間接的に出現する事になる．

しかしながら，どの値を基準値とすべきかは問題環境によって異なる．環境の中で最大の報酬獲得確率を持つ選択肢が 0.5 未満であるのならば，基準値を 0.5 から変更できない現在の *LS* の形式では永久的に探索行動が終わらない事を意味する．もちろん基本的な K 本腕バンディット問題は課題環境に関する情報を何も持たない，エージェントにとって未知なる環境として開始されるため，予め特定の基準値を知っていると想定すべきではない．しかしながら課題の試行錯誤によって，徐々に基準が生成されるようなアルゴリズムがあったとしても，基準値を変更できない現在の *LS* あるいは *LSK* では扱えない事になる．以下では問題点 1 の時と同様に，選択収束状態における収束の仕方を念頭に，基準値を変更可能にした *LS* の拡張モデルについて議論する．

4.2.4 改善点 2A：LS-VR

我々が考案する，基準値を変更可能に改良した *LS* を Loosely Symmetric model with Variable Reference (*LS-VR*) と名付ける．*LS-VR* が目的とすべき性質は以下の通りである．ここで変数 R は収束させたい任意の基準値である．

$$\lim_{S(a_{MT})\to 1.0} LSVR(e|a_{MT}; R) \approx S(e|a_{MT}) \tag{4.33}$$

$$\lim_{S(a_{MT})\to 1.0} LSVR(e|a_{LT}; R) \approx R \tag{4.34}$$

ここでバイアス項の議論に戻るが，選択収束状態にあるときよく知らない選択肢 a_{LT} に関する評価 LSK は式 4.19 から，以下の式で基準値に収束する．

$$\begin{aligned}
&\lim_{S(a_{LT})\to 0.0} LSK(e|a_{LT}) \\
=\ & \frac{w_L + b_K(\bar{e})}{w_L + l_L + b_K(\bar{e}) + b_K(e)} \\
=\ & \frac{w_L + l_L}{w_L + l_L + l_L + w_L} \\
=\ & \frac{(w_L + l_L)1}{(w_L + l_L)(1+1)} \\
=\ & \frac{1}{2}
\end{aligned} \tag{4.35}$$

ここで式 4.35 の 4 段目から，同一の選択肢に関する情報，$(w_L + l_L)$ で括られる箇所が 3 つある事がわかる．その 3 箇所にそれぞれパラメータ (χ_1, χ_2, χ_3) を付加すると式 4.36 となり，その $P(a_{LT}) \to 0.0$ に対する極限は式 4.37 である．ゆえに $LS\text{-}VR$ は式 4.37 が基準値 R に収束する事が $LS\text{-}VR$ を構成する最低限の必要条件になる．

$$\frac{\chi_1(w_i + b_K(\bar{e}))}{\chi_2(w_i + l_i) + \chi_3(b_K(\bar{e}) + b_K(e))} \tag{4.36}$$

$$\begin{aligned}
&\lim_{S(a_{LT})\to 0.0} \left(\frac{\chi_1(w_i + b_K(\bar{e}))}{\chi_2(w_i + l_i) + \chi_3(b_K(\bar{e}) + b_K(e))} \right) \\
=\ & \frac{(w_L + l_L)\chi_1}{(w_L + l_L)(\chi_2 + \chi_3)} \\
=\ & \frac{\chi_1}{\chi_2 + \chi_3}
\end{aligned} \tag{4.37}$$

逆によく知っている選択肢 a_{MT} に関する評価 LSK は式 4.18 から，以下の式で基準値に収束する．

$$\begin{aligned}
&\lim_{S(a_{MT})\to 1.0} LSK(e|a_{MT}) \\
=\ & \frac{w_H + b_K(\bar{e})}{w_H + l_H + b_K(\bar{e}) + b_K(e)} \\
=\ & \frac{w_H + l_L}{w_H + l_H + l_L + w_L} \\
=\ & S(E|A_H)
\end{aligned} \tag{4.38}$$

式 4.38 から，よく知っている選択肢 A_H に関する極限は式 4.35 のように，分子について同じ選択肢に関する変数で複数箇所を括る事が出来ない．即ち式 4.37 における χ_1 と χ_2 は変更不可である事がわかる $(\chi_1 = \chi_2 = 1)$．その条件において $\chi_3 = \rho$ を求めると式 4.40 になる．これらを式 4.36 に代入する事で $LS\text{-}VR$ は式 4.41 として定義できる．

$$R = \frac{\chi_1}{\chi_2 + \chi_3} = \frac{1}{1+\rho} \tag{4.39}$$

$$\rho = \frac{1}{R} - 1 \tag{4.40}$$

$$LSVR(e|a_i;R) = \frac{w_i + b_K(\bar{e})}{w_i + l_i + \rho(b_K(\bar{e}) + b_K(e))} \tag{4.41}$$

これによって式 4.33，4.34 は満たされる．また基準値 R の値は $0 < R \leq 1$ の範囲であれば任意に変更でき，その変更タイミングは問わず，試行錯誤の中で動的に変更する事も可能である．ここで式 4.42，4.43 を定義する事で式 4.11 と同様に *LS-VR* を重み付け平均として定義する事が出来る．

$$\omega_{VR} = \frac{R(w_i + l_i)}{R(w_i + l_i) + (1-R)(b_K(\bar{e}) + b_K(e))} \tag{4.42}$$

$$V_{VR}(\bar{e}) = \frac{b_K(\bar{e})}{b_K(\bar{e}) + b_K(e)} \tag{4.43}$$

$$\begin{aligned} & LSVR(e|a_i;R) \\ = & \frac{Rw_i + Rb_K(\bar{e})}{R(w_i + l_i) + (1-R)(b_K(\bar{e}) + b_K(e))} \\ = & \omega_{VR} S(e|a_i) + (1-\omega_{VR})\frac{R}{1-R} V_{VR}(\bar{e}) \end{aligned} \tag{4.44}$$

しかし，LS-VR にはある選択肢 a_i から得られた報酬の生起 (e) と不生起 ($\bar{e}$) の評価に対して排中律を満たさないという問題 (4.46) が存在する．これは LS-VR が主観確率モデルとしての LS の性質を失っている事を意味する．

$$1 = R + \overline{R} \tag{4.45}$$

$$1 \neq LSVR(e|a_i;R) + LSVR(\bar{e}|a_i;\overline{R}) \tag{4.46}$$

4.2.5 改善点 2B：LSX

上述した LS-VR は基準値を変更可能にしたが，報酬の生起 (e) と不生起 ($\bar{e}$) に対して排中律を満たさないという問題点があった．そのため，LS-VR と同様に基準値を変更可能にしながら，排中律を満たして主観確率モデルとしての LS の性質を保った拡張モデルである EXtended Loosely Symmetric model (LSX) を考案した．LSX は LS-VR と同じく式 4.47, 4.48 を満たす必要がある．また排中律を満たす事も目的とする (式 4.49).

$$\lim_{S(a_{MT})\to 1.0} LSX(e|a_{MT};R) \approx S(e|a_{MT}) \tag{4.47}$$

$$\lim_{S(a_{MT})\to 1.0} LSX(e|a_{LT};R) \approx R \tag{4.48}$$

$$1 = LSX(e|a_i;R) + LSX(\bar{e}|a_i;\overline{R}) \tag{4.49}$$

ここでバイアス項の議論に戻る．K 個の選択肢に対応したバイアス項 B_K を用いた LS 式である LSK は，LS における式 4.11 と同じように重み付き平均で記述する事ができる (式 4.55)．この式 4.55 と式 4.19 から，選択収束状態にあるときよく知らない選択肢 a_{LT} に関する評価 LSK は式 4.35 のような過程で基準値に収束する (背景化).

$$
\begin{aligned}
W_{B_K} &= B_K(\bar{e}) + B_K(e) && (4.50)\\
V_{B_K}(\bar{e}) &= \frac{B_K(\bar{e})}{W_{B_K}} && (4.51)\\
V_{B_K}(e) &= \frac{B_K(e)}{W_{B_K}} && (4.52)\\
1 &= V_{B_K}(e) + V_{B_K}(\bar{e}) && (4.53)\\
\omega^i_{B_K} &= \frac{S(a_i)}{S(a_i) + W_{B_K}} && (4.54)\\
LS(e|a_i) &= \omega^i_{B_K} S(e|a_i) + (1-\omega^i_{B_K}) V_{B_K}(\bar{e}) \\
&= \omega^i_{B_K} S(e|a_i) + (1-\omega^i_{B_K})(1 - V_{B_K}(e)) && (4.55)
\end{aligned}
$$

$$
\begin{aligned}
&\lim_{S(a_{MT})\to 1.0} W_{B_K} \\
=\ & S(\bar{e} \cup a_{LT}) + S(e \cup a_{LT}) \\
=\ & S(a_{LT}) && (4.56)\\
&\lim_{S(a_{MT})\to 1.0} V_{B_K}(e) \\
=\ & \frac{S(e \cup a_{LT})}{S(a_{LT})} \\
=\ & S(e|a_{LT}) && (4.57)\\
&\lim_{S(a_{MT})\to 1.0} LSK(e|a_{LT}) \\
=\ & \omega^{LT}_{B_K} S(e|a_{LT}) + (1-\omega^{LT}_{B_K})(1 - V_{B_K}(e)) \\
=\ & \frac{1}{2} S(e|a_{LT}) + (1-\frac{1}{2})(1 - S(e|a_{LT})) \\
=\ & \frac{1}{2}\big(S(e|a_{LT}) + 1 - S(e|a_{LT})\big) \\
=\ & \frac{1}{2}(1) && (4.58)
\end{aligned}
$$

それに対して鮮明化 (式 4.18) は，選択収束の定義上 $S(a_{MT}) \gg S(a_{LT})$ であるため，$\omega^{MT}_{B_K} = 1$ になる事によって行われる．

$$
\begin{aligned}
&\lim_{S(a_{MT})\to 1.0} LSK(e|a_{MT}) \\
=\ & \omega^{LT}_{B_K} S(e|a_{LT}) + (1-\omega^{LT}_{B_K})(1 - V_{B_K}(e)) \\
=\ & (1) S(e|a_{MT}) + (1-1)(1 - S(e|a_{LT})) \\
=\ & S(e|a_{MT}) && (4.59)
\end{aligned}
$$

重み付け変数である $\omega^i_{B_K}$ は $LSK(e|a_i)$，$LSK(\bar{e}|a_i)$ に対して不変であるため，これを書き換えてしまうと $1 = LSK(e|a_i) + LSK(\bar{e}|a_i)$ であるという排中律が満たされなくなる．そのため

式 4.48 を満たすためには，$V_{B_K}(\bar{e}) = \big(1 - V_{B_K}(e)\big)$ を一般化するべきだと考えられる．式 4.58 を見ると，客観的な評価値 $S(e|a_{LT})$ と，$V_{B_K}(\bar{(e)})$ の極限として現れる $-S(e|a_{LT})$ が打ち消しあって 1 が残り，それに $\omega_{B_K}^{LT}$ の極限である 1/2 が掛かって 0.5 に背景化しているのがわかる．$0.5 = R$ と置くと，$V_{B_K}(\bar{(e)})$ に含まれる 1 は $1 = 2R$ という意味を持つと捉える事ができる．よって任意の基準値 R への背景化 (式 4.48) を成す LSX は式 4.60 のようになると考えられる．

$$LS(e|a_i, R) = \omega_{B_K}^{i} S(e|a_i) + (1 - \omega_{B_K}^{i})(2R - V_{B_K}(e)) \tag{4.60}$$

鮮明化 (式 4.47) は LSK と同様に，$\omega_{B_K}^{MT} = 1$ になる事によって $S(e|a_{MT})$ のみが残る事によって実現される．これによって式 4.47，4.48 は満たされる．

また以下のように式をまとめる事によって，任意の選択肢 a_i に対する客観的な評価値 $S(e|a_i)$ と，$S(e|a_i)$ に対する 2 種の差分式 4.62，4.63 とを重み付け平均した価値関数と見なす事ができる．

$$\begin{aligned}
\omega_i &= 1 - \omega_{B_K}^{i} && (4.61)\\
\delta_i &= R - S(e|a_i) && (4.62)\\
\eta_i &= V_{B_K}(e) - S(e|a_i) && (4.63)\\
\beta &= \lim_{S(a_{MT}) \to 1.0} \frac{1}{\omega_i} \\
&= 2 && (4.64)\\
\bar{\beta} &= 1 - \beta \\
&= -1 && (4.65)\\
LS(e|a_i, R) &= S(e|a_i) + \omega_i(\beta\delta_i + \bar{\beta}\eta_i) && (4.66)
\end{aligned}$$

LSX は LS-VR と同様の性質を持ちながら，排中律を満たすよう LS を拡張したアルゴリズムである．すなわち LSX は LS の主観確率モデルとしての性質を保つため，LS-VR よりも正しい LS の拡張モデルであると考えられる．K 本腕バンディット問題に用いる価値関数としての性質はどちらも変わらないため，以降の議論では LS-VR の代わりに LSX を用いる．

4.3 オンライン LSX アルゴリズム

K 本腕バンディット問題における LSX の計算及び選択アルゴリズムを Algorithm 1 に示す．まず着目すべきは Algorithm 1 の 7〜11 行目においてバイアス項 ($b_K(e)$，$b_K(\bar{e})$，n_K) が予め計算されている点にある．バイアス項は地の不変性から全て選択肢に対して等しい値を持つため，それぞれの選択肢についての LSX の値を計算する度に計算し直す必要がない．毎 step に一度計算すれば良いという点で非常に節約的である．ここで定数 ϵ は正の微小な数を意味する．ϵ は変数 w_i，l_i の初期化に使用する変数であり，価値関数 LSX がゼロ除算を起こす事を防ぐために代入する．コンピュータシミュレーション上における ϵ は式 4.67 を満たす事が望ましいが，1.0 より充分に小さい値であれば成績に大きな影響を与えない．

$$1 \gg \epsilon \geq \sqrt{\text{倍精度浮動少数型実数の正の最小の値}} \tag{4.67}$$

Algorithm 1 オンライン LSX アルゴリズム

1: **while** $a_i \in A$ **do**
2: $w_i \leftarrow \epsilon$
3: $l_i \leftarrow \epsilon$
4: **end while**
5: $R \leftarrow 0.5$
6: **while** Current step $\leq$ Max step **do**
7: $a_{MT} \leftarrow \arg\max_{a_k} S(a_k)$
8: $a_{LT} \leftarrow \arg\min_{a_k} S(a_k)$
9: $b_K(\bar{e}) \leftarrow \frac{l_{MT} l_{LT}}{l_{MT}+l_{LT}}$
10: $b_K(e) \leftarrow \frac{w_{MT} w_{LT}}{w_{MT}+w_{LT}}$
11: $n_K \leftarrow b_K(\bar{e}) + b_K(e)$
12: **while** $a_i \in A$ **do**
13: $LSX(e|a_i, R) \leftarrow \frac{w_i + 2Rw_K - b_K(e)}{w_i + l_i + n_K}$
14: **end while**
15: $a_{Select} \leftarrow \arg\max_{a_k} LSX(e|a_k, R)$
16: 選択肢 a_{Select} を 試行して結果 e_{Gain} を得る
17: **while** $a_i \in A$ **do**
18: $w_i \leftarrow \gamma w_i$
19: $l_i \leftarrow \gamma l_i$
20: **end while**
21: **if** $e_{Gain} = e$ **then**
22: $w_{Select} \leftarrow w_{Select} + 1$
23: **else**
24: $l_{Select} \leftarrow l_{Select} + 1$
25: **end if**
26: $R \leftarrow R + \alpha(S(e|a_{Select}) - R)$
27: **end while**

Algorithm 1 の 18，19 行目と 26 行目に出現するパラメータ γ と α はそれぞれ忘却率と学習率と呼ぶ．忘却率 γ は 3.2 で表される変数 w_i，l_i の更新手法に用いられ，学習率 α は基準値 R の更新手法に用いられる．以下では観測量更新規則と基準値更新規則と，そこで用いられるパラメータ γ，α の機能的意味を説明する．

4.3.1 観測量更新法

Algorithm 1 の 17〜25 行目の記述は 3.2 で表される観測量の更新手順を表している．これらの手順を数式で記述すると以下の通りになる．

$$w_k = \begin{cases} \gamma w_k + 1 & (a_k = a_{Select} \wedge e) \\ \gamma w_k & (otherwise) \end{cases} \tag{4.68}$$

$$l_k = \begin{cases} \gamma l_k + 1 & (a_k = a_{Select} \wedge \bar{e}) \\ \gamma l_k & (otherwise) \end{cases} \tag{4.69}$$

ここでパラメータ γ は $0 \leq \gamma \leq 1$ の範囲を持つ．式 4.68，4.69 において $\gamma = 1$ ならば単なる頻度の更新と等しくなる．しかし $\gamma < 1$ だと，過去に得られた 全ての選択肢に対する情報である観測量 w_k，l_k は γ の分だけ圧縮されて過去の情報の重みが減衰していく．そのため γ は忘却率と呼ぶ事にする．例えば，問題の開始後 t step 時に選択肢 a_i を試行して結果が得られなかった $(\bar{e})$ とする．この場合には観測量 l_i に $+1$ されるが，t step 時の試行結果 $(a_i \cap \bar{e})$ という情報は $t + m$ step 後，γ^m だけ圧縮された量として扱われる．このような忘却率を用いた更新規則をとるため厳密には頻度とは言えない．これが変数 w_k，l_k を頻度と呼ばず観測量と呼ぶ理由である．しかしながらこの更新規則には古い情報を圧縮するという利点が存在する．まず t step 時の観測量の合計値を n_t と定義する (式 4.70)．この更新規則において $\gamma < 1$ であるとき，充分に試行回数を重ねれば更新前と更新後の観測量の合計値がほぼ等しくなる $(n_t = n_{t+1} = n_{max})$．これを観測量に対する平衡状態とする．頻度情報の合計値 n_t に対する更新量は step 毎に $+1$ であるため，平衡状態での n_{max} は式 4.72 で定義できる．

$$n = \sum_k (w_k + l_k) \tag{4.70}$$

$$n_{max} = \gamma n_{max} + 1 \tag{4.71}$$

$$n_{max} = \frac{1}{1-\gamma} \tag{4.72}$$

この事から加えられる更新量 $+1$ は観測量の合計値の最大値 n_{max} に対して，式 4.73 の重みを持つ事になる．

$$\frac{1}{n_t} \geq \frac{1}{n_{max}} = \frac{1}{\frac{1}{1-\gamma}} = 1 - \gamma \tag{4.73}$$

具体的には，例えば $\gamma = 0.999$ である場合，観測量の合計値 n_t と比較した更新量は最低でも 0.1% の割合の重みを持つ事を意味する．ここで観測量の合計値 n_t は全体の記憶量であり，n_{max} は全体の記憶量の上限と言える．故に上述した重みとは，記憶の全体量に対する新しい試行結果の影響度と言い換える事が出来る．このように忘却率 γ を用いた観測量の更新規則には，古い情報を圧縮して新しい情報の影響度の下限を保つ役割がある．$\gamma = 1$ である場合，式 4.72 から，観測量の合計値の上限 n_{max} は無限大に発散してしまい，平衡状態が発生しない．そのため全体の記憶量と比較した新しい情報の影響度 (式 4.73) も下がり続け，いずれ影響度が 0 になる．これは全ての選択肢に対する真の報酬確率が常に一定である場合ならば問題無い．しかしながら，現実の環境は不安定で非定常であると考えるべきである．そのため，このような新しい情報に対する影響度の下限を保つ仕組みは，より現実的な非定常バンディット問題において有効であると考えられる．

4.3.2 漸進的基準値更新法

Algorithm 1 の 26 行目の記述は LSX において変更可能な基準値 R の更新手順を表している．これを数式で記述すると以下の漸化式になる．

$$R = R + \alpha(S(e|a_{Select}) - R) \tag{4.74}$$

前述した通り R とは満足化において，探索を中止するための基準値である．LSX では選択肢の中に試行錯誤によって観測された報酬獲得割合 $S(E|A_i)$ が基準値 R を越えなければ，間接的な探索行動が行われ，そう出なければ利益追求が行われる．基準値 R は端的に言えば，観測情報から探索行動と利益追求行動のどちらを行うか決定する境界線と見なす事が出来る．この指標は問題当初こそ暫定的で経験的に得られた曖昧な値 (ここでは $R = 0.5$) であるが，試行錯誤を繰り返すうちに変化して行くと考えられる．例えば，報酬を多く獲得すれば，その分だけ期待する報酬の獲得率は上昇する．逆に報酬が得られない状況が続けば，期待は損なわれていく．式 4.74 は正にそのような期待の上昇と下降という形で基準値の変化を表現している．ここで α は更新量の影響度合いを定めており，強化学習における価値関数の更新手法に習って学習率と呼ぶ．厳密に言えば，式 4.75 のように，報酬が得られた時 ($e_{Gain} = e$)，得られなかった時 ($e_{Gain} = \bar{e}$) を正確に反映するために，その step 時に得られた 1 or 0 の報酬との差分を更新に用いるべきだが，確率の範囲において 0 と 1 の差は大きく，よく振動してしまう．しかしながら振動を抑えるために学習率 α を小さい値にすると学習が遅くなり，探索の誘発が不安定になる．そのため，本アルゴリズムの更新規則では安定のために $S(e|a_{Select})$ と R の差分値を用いた．これを漸進的基準値更新法と呼ぶ．

$$R = R + \begin{cases} \alpha(1 - R) & (e_{Gain} = e) \\ \alpha(0 - R) & (e_{Gain} = \bar{e}) \end{cases} \tag{4.75}$$

ここで選択肢 a_i の真の報酬獲得確率を P_i とする．そして P_i が 1 番目に高い選択肢の真の報酬獲得確率を P_{First}，2 番目に高い確率を P_{Second} とした時，満足化から，R が式 4.76 の条件を満たせば，最も高い報酬獲得確率を持つ選択肢 A_{First} を観測するまで探索し続ける事が可能になり，結果的に必ず最も良い選択肢を見つけ出す事が可能になる．

$$P_{Second} < R < P_{First} \tag{4.76}$$

現時点で式 4.74 は最適な条件である式 4.76 を必ず満たすような更新規則ではない．しかしながら，式 4.74 の更新規則であるが故のメリットも存在する．第一に選択肢であるバンディットマシンの背景にある構造や確率分布を想定しなくて使える事が挙げられる。第二に式 4.74 は強化学習の更新規則を意識して考案したため，確率の値域に限定された規則ではない事が挙げられる．そして $a_{Select} = a_{First}$ であれば，式 4.74 は限りなく式 4.76 の条件に限りなく当てはまるようになる．これらの性質から前述した非定常な K 本腕バンディット問題においても良い更新規則なのではないかと考えられる．

4.3.3　統計的基準値更新法

前述した漸進的基準値更新法は，環境の知識がまったくない場合に用いる消極的な手法であった．しかし報酬の出現がベルヌーイ試行であり，かつ真の期待値が不変である K 本腕バンディット問題では，統計的な知見により試行結果から真の報酬獲得確率を推定する事ができる．そのため P_{First} と P_{Second} の楽観的な推定という形で基準値 R_{CC} を定義することもできる (式 4.81)[Kohno 15]．ここで $U(e; a_i)$ 関数は i 番目の選択肢に関する 結果 e についての標本平均 $S(e|a_i)$ から，母平均 $P(e|a_i)$ を区間推定する際の上限値を定義する関数である．$c = 2.58$ である場合，$P(E|A_i) + c\sigma_i/\sqrt{n_i}$ は選択肢 A_i が結果 E を生起させる確率の標本平均の 95% 信頼区間の上限に近似する．試行結果が充分に蓄えられれば，標本平均の 95% 信頼区間の上限は

$P(E|A_i)$ に収束するため，最終的に式 4.81 は最適な基準値の条件式 4.76 を満たす．我々はこの基準値 R_{CC} を LSX 価値関数に用いたアルゴリズムを統計的 LSX アルゴリズムと名付けた．

$$U(e;a_i) = S(e|a_i) + c\sqrt{\frac{V_u(e;a_i)}{N(a_i)}} \tag{4.77}$$

$$a_{upper} = \arg\max_{a_j \in A} U(e;a_j) \tag{4.78}$$

$$a_{lower} = \arg\max_{a_j \in (A \cap \overline{a_{upper}})} U(e;a_j) \tag{4.79}$$

$$\tag{4.80}$$

$$R_{CC} = \frac{U(e;a_{upper}) + U(e;a_{lower})}{2} \tag{4.81}$$

しかしながら，これらはあくまで真の報酬獲得確率が不変である定常環境において用いれるのみで，選択に対して報酬の出現がベルヌーイ試行でなかったり，途中で報酬獲得確率が変化しないという保証がない場合には用いる事ができない．

4.3.4 探索の仕組み

LSX アルゴリズムには大きく分けて三つの探索的効果がある．ここでいうスイッチングとは直前のステップで選んだ選択肢とは別の選択肢を選ぶ事であり，探索と関連がある行為である．

- 満足化スイッチング
- バウンススイッチング
- 撹拌スイッチング

満足化スイッチングとは，前述した満足化の効果から，基準を満たす選択肢を見つけるまで探索する事である．観測された報酬確率が基準を満たす選択肢を見つけたら満足化に伴う探索は終了する．

バウンススイッチングとは信頼性考慮と競合によって引き起こされる探索である．エージェントの選択が選択収束であり，$a_j \in A \cap (\overline{a_{MT}} \cap \overline{a_{LT}})$ かつ $S(a_j) \approx S(a_{LT})$ である場合，選択肢 a_j の LSX による評価は式 4.82 の通りになる．以上の条件下で発生する “背景化” に似た現象を “準背景化” と呼ぶ．

$$\begin{aligned} & \lim_{S(a_j)\to S(a_{LT})} LSX(e|a_j,R) \\ \approx\ & \omega^j_{B_K} S(e|a_j) + (1-\omega^j_{B_K})(2R - S(e|a_{LT})) \\ \approx\ & \frac{1}{2}S(e|a_j) + \frac{1}{2}(2R - S(e|a_{LT})) \\ \approx\ & R - \frac{1}{2}(S(e|a_j) - S(e|a_{LT})) \end{aligned} \tag{4.82}$$

準背景化は背景化 (式 4.48) と異なり，$S(e|a_j) < S(e|a_{LT})$ なら基準値 R を下回り，$P(e|a_j) > P(e|a_{LT})$ なら基準値 R を上回る．基準値を上回る事で基準値 R が最適な条件である式 4.76 を満たしていても，a_j が a_{MT} の評価を越えてしまう場合がある．これを我々は a_j に関する “バウンス” と呼ぶ．LSX アルゴリズムは各選択肢の評価値に対して常に greedy に選択を行うた

め，バウンスした a_j の中で最も高い選択肢が選択される．我々はこのような選択肢収束の阻害をバウンススイッチングと呼ぶことにした．このような選択がされる事で選択肢 a_j に関する選択割合 $S(a_j)$ が最も選択されていない選択肢の割合 $S(a_{LT})$ より充分大きくなる事で，準背景化の条件 ($S(a_j) \approx S(a_{LT})$) から外れてバウンスが解除される．このように準背景化によって $S(e|a_j) > S(e|a_{LT})$ となる全ての a_j において $S(a_j) \gg S(a_{LT})$ が満たされてバウンスが解除されるまでバウンススイッチングによって選択収束は阻害され続ける．そして満足化スイッチングとバウンススイッチングが終わると選択収束していく．この状態を “安定状態” と呼ぶ事にした．

最後に，撹拌スイッチングとは安定状態や選択収束に関係なく発生するスイッチングであり，それは更新による基準値 R の揺らぎから生じる．LSX アルゴリズムでは基準値 R は式 4.74 から常に値が振動している．また，任意の選択肢の客観的価値 $S(e|a_i)$ も報酬が 0 or 1 であるため振動している．これにより偶然に安定状態における $S(e|a_{MT})$ が基準値 R を下回る事がある．すると背景化して R に近似している最も選択されていない選択肢 a_{LT} の LSX における評価値が上回り，探索的に選択される ($a_{Select} = a_{LT}$)．しかし選択された事によって $S(a_{LT})$ が上昇するため，a_{Select} は式 4.17 に示す最も選択されていない選択肢 a_{LT} の条件から外れてしまう場合がある．すると新たに別の選択肢が a_{LT} になるため，それにともなってバイアス項である式 4.22, 4.23 が step の移行に対して不連続に変化してしまう．これによって安定状態が解除され，再び擬似的なバウンススイッチングが頻発するようになる事がある．このように選択によって最も選択されていない選択肢 a_{LT} が変更される事を “撹拌” と呼び，それに伴う擬似的なバウンススイッチングの群発を撹拌スイッチングと呼ぶ事にした．擬似的なバウンススイッチングが収まるとまた安定状態になる．しかし再び確率の揺らぎや環境の変化によって撹拌が起こると撹拌スイッチングが引き起こされる．これらの性質が LSX は探索的な行動を促すであろうと想定される．

第5章　K本腕バンディット問題シミュレーション

オンライン *LSX* アルゴリズムの性能を調べるため，K 本腕バンディット問題においてシミュレーションを行い，その結果から他のアルゴリズムとの比較を行った．シミュレーションの環境設定は 3 種類を用意した．まずは複数の選択肢が始めから有している報酬が生起する真の確率 (真の報酬生起確率) が一定である場合を想定した．これを定常環境と呼び，各アルゴリズムに対する *LSX* の性質を考察した．次に同期的な非定常環境での成績を比較した．同期的な非定常環境とは一定 step 毎に全ての選択肢の報酬生起確率が一斉に再設定される環境を意味した．この課題では環境の不安定さを表し，それに対する適応度合いを検証した．次に非同期的な非定常環境での成績を比較した．非同期的な非定常環境とは一定確率で非同期にそれぞれの選択肢の報酬確率が再設定される環境を意味する．単純な成績の比較は同期的環境の方がしやすいものの，選択肢の変化が同期的であるのは不自然であるため，より自然な非定常環境を想定した場合でも成績を比較した．

5.1　共通設定と指標

本シミュレーションに選択を行うエージェントは，選択アルゴリズムに従い決定された選択肢を実際に試行して，報酬の生起 (e)・不生起 ($\bar{e}$) を観測するまでを 1 step とする．各指標は決められた step 数 (総 step 数) を 1 simulation とし，それを 1,000 回 (1,000 simulations) 行って算出された指標を平均して提示される．各指標のプロット点の数は 1,000 プロットであり，1 プロットの持つ値は (総 step 数 /1,000) の間に得られた指標の平均である．選択肢の数は 20 arms であり，それぞれシミュレーション開始時に一様分布から真の報酬生起確率が独立に設定される．

本シミュレーションにおいて指標は 3 種類使用する (正解率，後悔の度合い，入替率)．それぞれの指標の定義について以下に詳しく説明する．

5.1.1　正解率

正解率とは全ての選択肢の中で最も高い報酬生起確率を有する選択肢を正解である選択肢としたとき，各 step において正解の選択肢を選択できていたかの指標であり，正解の選択肢を選択できた割合によって定義される．正解率は 100% に近づく程良い成績であるといえ，制限時間内で正解率がどの程度まで上昇するかが端的な正確さを表す指標であると考えられる．

5.1.2　後悔の度合い

後悔の度合いとは，上記で定義した正解の選択肢をシミュレーション開始時から現時点までずっと選択した場合に得られる理想的な報酬期待値と，実際に獲得された報酬との差によって

定義される．数式では以下で定義される．ここで t は現時点の step 数，P_i は i 番目の選択肢の真の報酬生起確率，$r(t)$ は t step において獲得した報酬値 (生起であれば 1 ，不生起であれば 0) を意味する．

$$regret(t) = \sum_{k=1}^{t} \big(\max(P_i) - r(k) \big) \tag{5.1}$$

後悔の度合いにおいてその値は低い方が好ましい．ある step の後悔の度合いの値はそれまでの損失の蓄積であるため，それまでの選択がその時点での報酬の最大化に対してどの程度悪い選択だったかを示す．ここで重要なのは，正確さを重視するなら多くの探索的な選択を行わなければならないという速さと正確さのトレードオフに対してどのような意義を有する指標であるか，である．もし真に正しい選択肢を発見している状態で，そうと知らずに探索を行えば，その分だけ報酬を損なう事になる．しかも選択肢が 3 つ以上あるなら，探索にもより効率が求められ，あまりにも報酬生起確率が低いと考えられる選択肢より，ある程度高い選択肢を優先して探索する等の性質が求められる．即ち後悔の度合いの情報には探索の効率が含められる．ただしそれが全てではなく，どれだけ選択が正しさに近づいたかの過程とそのために犠牲として払った損失が時間的効率的に畳み込まれている．そのため詳細までわかる指標ではないものの，速さと正確さのトレードオフに対する包括的な指標と見なす事ができる．

5.1.3　入替率

入替率とは，1 step 前に選択した選択肢から，現在の step で選択した選択肢がどれだけ変更されたかの割合である．シミュレーション 1 回の 1 系列データでは 0 と 1 の二値で示されるバイナリ行列だが，1,000 回のシミュレーション結果を平均する事で割合として示される．短期的な報酬獲得 (速さ) を重視するならば，この選択率はなるべく少なくなるべきであるが，報酬の最大化という目的に対して直接的な指標であるとは言えない．特に非定常環境では選択率が低く固定されてしまう事は望ましく無く，環境の変化に応じて柔軟に変化すべきであり，その指標として扱う事も出来る．

5.2　比較に用いたアルゴリズム

N 本腕バンディット問題における LSX の性質を示すため，幾つかのアルゴリズムを比較に用いた．具体的には，従来通りの LS や，本研究において LSX を考案するための過程として考案された基準を更新しない一般化 LS モデル (LSK) によって選択肢を評価するアルゴリズム，現在バンディット問題で非常に良い成績を有しており，また統計的背景から考案されたアルゴリズムである UCB1-tuned，更に理想的な基準値 R_{OPT} を常に知っている場合の 最適基準 LSX アルゴリズムと比較した．各アルゴリズムでは上述した過去の情報を圧縮して更新する，忘却率 γ を $\gamma = 1.0$(非圧縮)，$\gamma = 0.999$ の二通りでシミュレーションした．また，オンライン LSX アルゴリズムの基準値 R の学習率 α は $\alpha = 0.1$ とした．以下に比較に用いたアルゴリズムに関する説明と目的を示す．また 2 本腕バンディット問題において価値関数に LS を用いた選択アルゴリズムは，2 通りを越える選択肢を持つ N 本腕バンディット問題に対応したアルゴリズムではないため，本シミュレーションでは比較に用いない．

5.2.1 *LSK* を価値関数とした選択アルゴリズム

K 本腕バンディット問題において価値関数に *LSK*(式 4.32) を用いた選択アルゴリズム．これを便宜上 *LSK* アルゴリズムと呼ぶ．*LSK* は *LSX* を汎化したモデルであるが，逆に *LSK* は R の更新を行わず，初期値 $R = 0.5$ に固定された特殊な *LSX* モデルであるとも考えられる．アルゴリズムに関しても Algorithm 1 から R の更新に関する機能を省き，R を常に初期値に固定する事で実現される．このアルゴリズムとの比較は，基準値 R を学習する事が選択や成績にどのような性質をもたらすかを比較によって表す事を目的としている．本章で示される実験結果においては *LSK* または *LSK* γ という名称表され，単に *LSK* ならば $\gamma = 1.0$(非圧縮) であり，*LSK* γ は $\gamma = 0.999$ の場合を示す．

5.2.2 UCB1-tuned

UCB1-tuned とは K 本腕バンディット問題において最も優れているとされているアルゴリズムの一種である [Wang 05]．UCB1-tuned は式 5.2 によって定義される価値関数において最も高く評価される選択肢を選択するアルゴリズムである．ここで変数 n_i は選択肢 a_i を試行した回数，$r_{k,i}$ は選択肢 a_i を k 回目に試行したときの報酬 (1 or 0) を意味する．

$$P(E|A_i) + \sqrt{\frac{\ln n}{n_i} min\left(1/4, V_i(n_i)\right)} \tag{5.2}$$

$$V_i(s) = \frac{1}{s}\sum_{k=1}^{s} r_{k,i}^2 - P(E|A_i)^2 + \sqrt{\frac{\ln n}{s}} \tag{5.3}$$

このアルゴリズムとの比較の目的は，既存の優れたアルゴリズムとの成績比較にある．また，非定常環境という複雑な環境においての汎用性についても比較から示す事を目的としている．本章で示される実験結果においては *UCB1-tuned* または *UCB1-tuned* γ という名称表され，単に *UCB1-tuned* ならば $\gamma = 1.0$(非圧縮) であり，*UCB1-tuned* γ は $\gamma = 0.999$ の場合を示す．

5.2.3 最適基準 *LSX* アルゴリズム

どれが高い報酬獲得確率を持つ選択肢 A_{First} なのかは未知であるが，1 番高い報酬獲得確率 P_{First} と 2 番目に高い報酬獲得確率 P_{Second} の値を常に正確にしる事が出来るという前提のもと，式 5.4 によって基準値 R を統計的知識から決定する．そのように基準値を推定する点を除いて，オンライン *LSX* アルゴリズムと同様の選択過程を持つアルゴリズムを最適基準 *LSX* アルゴリズムと名付ける．本章で示される実験結果においては LSX_{opt} または LSX_{opt} γ という名称表され，単に LSX_{opt} ならば $\gamma = 1.0$(非圧縮) であり，LSX_{opt} γ は $\gamma = 0.999$ の場合を示す．

$$R_{opt} = \frac{P_{Second} + P_{First}}{2} \tag{5.4}$$

通常の K 本腕バンディット問題の枠組みではエージェントは P_{First}，P_{Second} を常に正確に把握する事は不可能であるため，このアルゴリズムは K 本腕バンディット問題に対する実用的なアルゴリズムであるとは言えない．飽くまでも本シミュレーションでは理想的な場合の *LSX* の性質を示す事とオンライン *LSX* や *UCB1-tuned* との比較を目的として最適基準 *LSX* アルゴリズムを扱う．また，将来的に基準値 R をより適切に計算可能となった場合に表れる *LSX* アルゴリズム のポテンシャルを示す事を目的としている．

5.2.4　統計的基準 LSX アルゴリズム

通常のオンライン LSX アルゴリズムで用いられる漸進的基準値更新法ではなく，定常環境でより正確な基準値の動的な獲得が可能な統計的基準値更新法を用いたアルゴリズムを比較に用いる．基準値の更新に統計的基準値更新法を用いる以外，オンライン LSX アルゴリズムと同様の選択過程を持つアルゴリズムを統計的基準 LSX アルゴリズムと名付ける．本章で示される実験結果においては LSX_{cc} または LSX_{cc} γ という名称表され，単に LSX_{opt} ならば $\gamma = 1.0$(非圧縮) であり，LSX_{cc} γ は $\gamma = 0.999$ の場合を示す．本アルゴリズムは定常環境における動的な報酬の獲得手法の一例である．しかし非定常環境では統計的な推定の前提が成り立たないため，非定常環境でのシミュレーションでは扱わない．

5.2.5　メタバンディットアルゴリズム

非定常な K 本腕バンディット問題に対しては一般的に メタバンディットアルゴリズムが用いられる [Hartland et al. 06]．メタバンディットとは環境の変化の検出と，検出後の処理の組み合わせからなる．環境の変化は下記の Page-Hinkley 統計量 PH_T から現在最も高い報酬が得られている選択肢から報酬獲得割合の変化から検出する (PH_T が閾値 δ を超えると環境が変化したと判断する).

$$\overline{r_t} = \frac{1}{t}\sum_{l=1}^{t} r_l \tag{5.5}$$

$$m_T = \sum_{t=1}^{T} (r_t - \overline{r_t} + \delta) \tag{5.6}$$

$$M_T = \max\{m_t, t = 1...T\} \tag{5.7}$$

$$PH_T = M_T - m_T \tag{5.8}$$

$$Change\ alarm = \begin{cases} true & (PH_T > \lambda) \\ false & (otherwise) \end{cases} \tag{5.9}$$

そして変化が検出されると “従来の観測情報を引き継いだままの旧エージェント” に対して “観測情報を初期 step の状態に戻した新規エージェント” を生成する，その後 L step の間，旧エージェントと新規エージェントのどちらで意思決定を行うか，を二つの選択肢として，実際に環境のどの選択肢を試行するかの前に，上位エージェントが意思決定を行う．この L step の間をメタバンディット期間と呼ぶ．そしてメタバンディット期間終了後，旧エージェントと新規エージェントで高い報酬を得られていた方を残し，一方のエージェントと上位エージェントを破棄する．その後また，環境の変化の検出を再開する．

メタバンディット期間に実際の環境選択肢を試行して得られた報酬情報は新旧エージェントに共有される．両者の違いはそれ以前の観測情報を持っているか否かである．また，上位エージェントにも新旧どちらのエージェントが意思決定を担い，得られた報酬情報なのかという情報が保存されていく．また新旧エージェント，上位エージェントで行われる全て意思決定は UCB1-tuned アルゴリズムによって行われる．

上記でも触れたとおり，メタバンディットアルゴリズムは非定常 K 本腕バンディット問題では一般的に使われているアルゴリズムである．しかし対応できる非定常環境は限定されている．本来，検出すべき K 本腕バンディット問題の環境の変化には，

- 最適な選択肢の報酬確率が変化し他の選択肢の報酬確率以下に減少する
- 他の選択肢の報酬確率が変化し最適な選択肢の報酬確率以上に上昇する
- 上記が両方起こり逆転する

がある．最適な選択肢が変化して減少したならば，メタバンディットアルゴリズムで検出できるが，その他二つに関してはメタバンディットアルゴリズムでは検出できない [Hartland et al. 06]. この点も考慮して，非定常 K 本腕バンディット問題では割引率を用いたオンライン LSX アルゴリズムとメタバンディットアルゴリズムの比較を行った．本章で示されるシミュレーションでは非定常バンディット問題においてのみ用いられ，実験結果においては *Meta UCB1-tuned* という名称表される．

5.3 シミュレーション 1：定常

前述した共通設定の通り，本シミュレーションに選択を行うエージェントは，選択アルゴリズムに従い選択を総 step 数 として 500,000 steps 行い，その 1,000 simulations 分を平均して各種指標を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，シミュレーション開始時に一様分布から独立に決定されて以降 500,000 steps の間は変化しない．1,000 simulations のあいだ毎回，全ての選択肢の真の報酬生起確率はサンプリングしなおされる．比較するアルゴリズムには，LSK，LSX，LSX_{cc}，LSX_{opt}，$UCB1\text{-}tuned$ (全て $\gamma = 1.0$ と 0.999 の両方について，$\gamma = 0.999$ を用いた場合はモデル名 γ と表記する) を用いた．前述したメタバンディットアルゴリズムは非定常バンディット問題のためのアルゴリズムであるため定常バンディット問題シミュレーションからは除外した．

5.3.1 シミュレーション 1 の結果

以下に定常環境におけるシミュレーション 1 の結果を示す．図 5.1 は正解率の推移であり，図 5.2 は後悔の度合いの推移であり，図 5.3 は入替率の推移である．

まず，20 通りの選択肢の中から，報酬獲得確率が真に最も高い選択肢を選択できた割合である “正解率” に関しては 1.0 = 100% に近づく程に良い成績であると言える．大まかには LSK は 20 通りの選択肢に関しては正解率に関して良い成績は得られなかった．それに対して，LSX, LSX_{opt}，$UCB1\text{-}tuned(\gamma = 1.0)$ は最終的に 0.8 を越える成績を残した (表 5.1)．正解率，後悔の度合い共に非圧縮の最適基準 LSX アルゴリズム (LSX_{opt}) が最も良い成績を示した (正解率 0.997)．次いでベルヌーイ的バンディット問題の統計的性質を元に，動的に基準を推定する統計的基準 LSX_{cc} アルゴリズム (LSX_c) が良い成績を示した．これは LSX において直感的パラメータである基準値 R が適切である際の潜在的な性能を示している．しかしながら，前述の二つの LSX アルゴリズムに比べて応用範囲は広いものの，最適化のための背景を持たない基準値 R を漸進的に更新するオンライン LSX アルゴリズム (LSX) でも約 5,000step ～ 7,500 step までは過去の情報の圧縮，比圧縮に限らず，非圧縮の UCB-tuned より高い成績を示す．これは統計的な推定量を用いて長期的な正確さのみを重視する $UCB\text{-}tuned$ との方略の違いを端的に表していると考えられる．また忘却率 $\gamma = 0.999$ である場合の $UCB1\text{-}tuned$ γ は $\gamma = 1.0$ である非圧縮 $UCB1\text{-}tuned$ に比べて著しく成績が損なわれている事がわかる．これは UCB1-tuned で扱う価値 (式 5.2) が統計的な仮説に基づく繊細な数値であるため，過去の情報を圧縮するという更新方式に対応できなかったためである．忘却率 γ を用いた更新への対応という点に関し

ては，最適基準 LSX アルゴリズム (LSX_{opt}) や統計的基準 LSX_{cc} アルゴリズム (LSX_c) も巧く対応しているとは言い難い．それに対して，オンライン LSX アルゴリズム (LSX) は正解率において，むしろ圧縮 ($\gamma = 0.999$) した方が非圧縮 ($\gamma = 1.0$) の場合と比べて高い成績を有している．これは $\gamma = 0.999$ によって過去の情報を圧縮して更新される事で，前述した “撹拌” や “バウンス” が発生し易くなる事に起因すると考えられる．これは図 5.3 にもあらわれており，LSX γ ($\gamma = 0.999$) は入替率が減少し切っていない．前述した通り，LSX が入替率に表れる探索行動を完全に行わなくなるためには，(1) 基準点より高い報酬生起確率を持つ選択肢を見つけている事と，(2) バウンスが起こらない安定状態になっている事，(3) 基準点と現在選択している選択肢の報酬獲得確率の間に充分な差があり，“撹拌” が起こりえない事が挙げられる．少なくとも $\gamma = 0.999$ は過去の情報を圧縮する事で，特に (3) の状態を満たし難くなる．しかしながら環境が定常的である保証が無い場合は，探索行動を完全にしなくなる事は良い性質とは言い難い．そのような非定常環境につては次項以降でのシミュレーションで議論する．ここで特筆すべきは図 5.1 に示した正解率の推移での傾向のみならず，図 5.2 に示す後悔の度合いの推移でも，$LS - VR$ が過去の情報の圧縮 ($\gamma = 0.999$) に対して他に無い傾向が見られる事である．最も理想的に選択した際に得られる報酬の期待値に対する累積した損失の大きさを示す後悔の度合いでは，値はなるべく小さい方が好ましい．更に言えば，その上昇の仕方は線形的でなく，対数的に収束する事が望まれる．UCB1-tuned は対数的な収束をする事が知られており，本シミュレーションでも同様の結果を示している．それに対して，本研究で我々が提案する LSX は基準値が最適であるならば UCB1-tuned よりも損失か少なく，それでいて速く対数的な収束を示した．この結果から K 本腕バンディット問題において優れた性質を示している．しかしここで注目すべきは LSX(非圧縮) と LSX γ(圧縮) における成績の差である．正解率を表す図 5.1 において LSX γ(圧縮) はこの LSX(非圧縮) に勝っている．しかしながら後悔の度合いである図 5.2 では成績が逆転している．情報更新の際に $\gamma = 0.999$ を用いて古い情報を圧縮し，新しい情報に対する重みの最低保証を導入すると，基本的には正解率，後悔の度合いという両指標において著しい成績の減少が見られる．しかしながら，LSX においてのみは，後悔の度合いが悪化する代わりに，正解率ではむしろ 10% 程の上昇が見られた．この差は図 5.3 で得られた情報を圧縮することによる探索率の下げ止まりに起因すると考えられる．ここで最適な基準を持つはずの 最適基準 LSX アルゴリズム (LSX_{opt}) やそれを動的に獲得する 統計的基準 LSX アルゴリズム (LSX_{cc}) では圧縮すると成績が損なわれるのかという疑問が生じる．その答えは，オンライン LSX アルゴリズムにおける非圧縮時と圧縮時の入替率の違いから推測できる．非圧縮の LSX の入替率がほぼ 0% に収束しているのに対し，圧縮している LSX($\gamma = 0.999$) の入替率は 1.3% 程度に留まっている．つまり，圧縮しない LSX は探索を完全に止めての正解率であり，本シミュレーションの設定では結果的に正しい選択肢を選べる割合が 82.1% であると見なせる．その分非常に速く選択肢を見つけ出している．対して圧縮した LSX($\gamma = 0.999$) は探索し続けている場合の結果であり，後悔の度合いに表れる結果はそれぞれ別の要因から来ている．前述した通り，基本的に情報の圧縮は価値の推定を不安定にさせるため，成績には悪影響を及ぼす．最適基準 LSX アルゴリズム (LSX_{opt}) やそ計的基準 LSX アルゴリズム (LSX_{cc}) に関しても同様の事が言えるため，オンライン LSX と異なり，圧縮によって成績が低下したのだと考えられる．

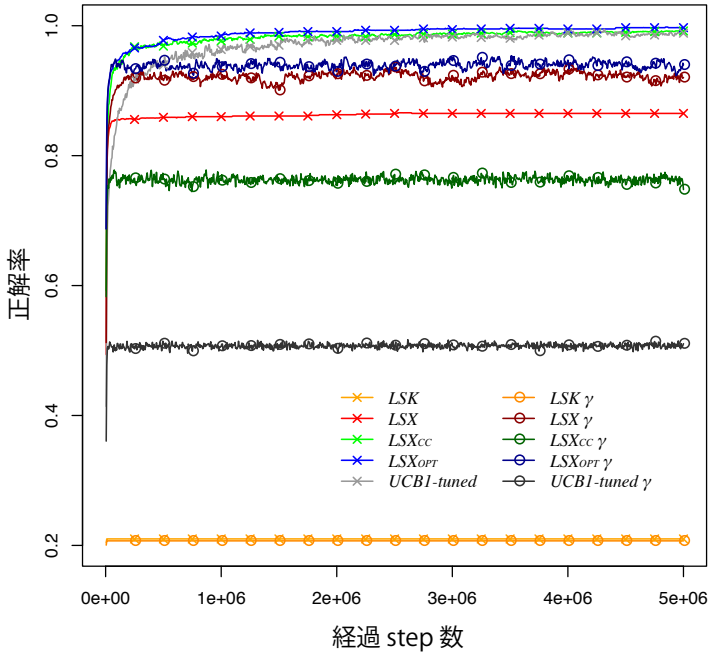

図 5.1: 定常 20 本腕バンディット問題：正解率

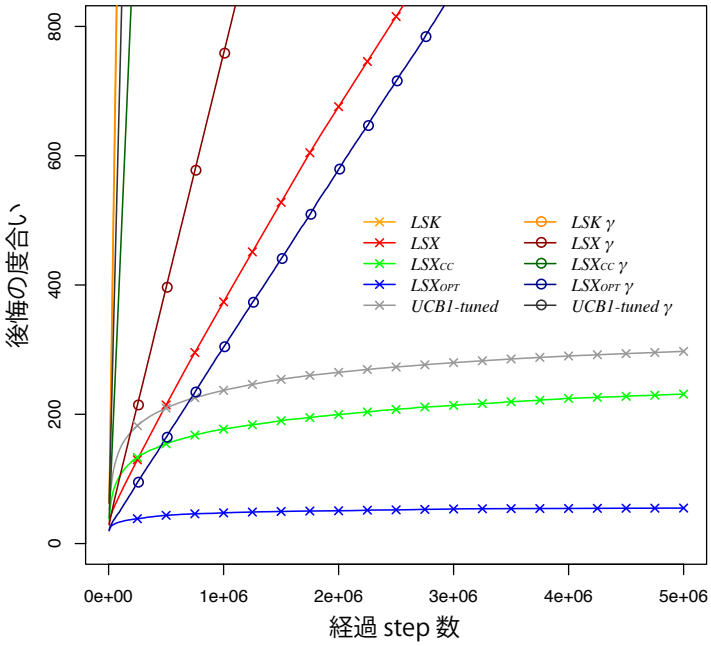

図 5.2: 定常 20 本腕バンディット問題：後悔の度合い

5.4 シミュレーション 2A：非定常-同期

前述した共通設定の通り，本シミュレーションに選択を行うエージェントは，選択アルゴリズムに従い選択を 総 step 数 として 100,000 steps 行い，そのシミュレーション 1,000 回分を平

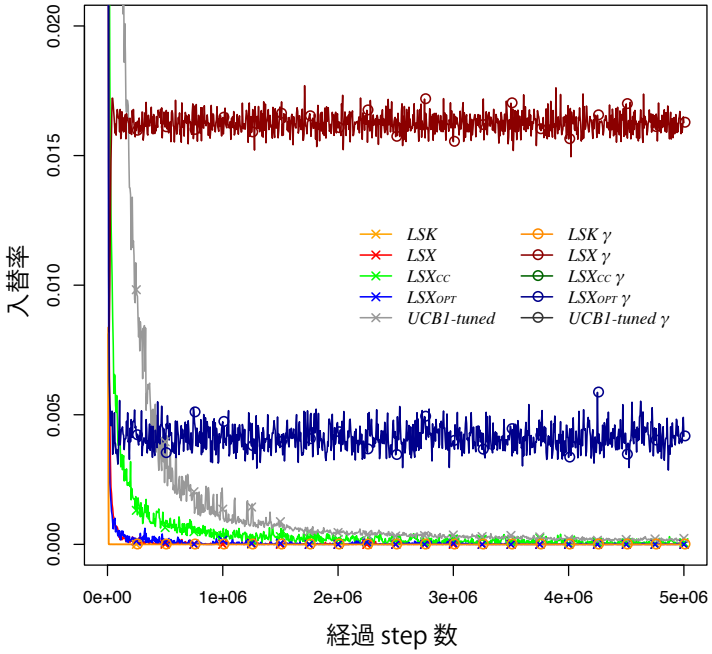

図 5.3: 定常 20 本腕バンディット問題：入替率

表 5.1: 定常 20 本腕バンディット問題：499,501 ～ 500,000 step の指標平均値

	正解率	後悔の度合い
LSK	0.210	61503.3
$LSK\ \gamma$	0.207	57290.6
LSX	0.865	1503.3
$LSX\ \gamma$	0.920	3660.7
LSX_{cc}	0.992	231.2
$LSX_{cc}\ \gamma$	0.748	20762.6
LSX_{opt}	0.997	54.9
$LSX_{opt}\ \gamma$	0.940	1401.7
$UCB1\text{-}tuned$	0.989	297.3
$UCB1\text{-}tuned\ \ \gamma$	0.510	35125.0

均して各種指標を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，シミュレーション開始時に一様分布から独立に決定される．前述したシミュレーション 1 の定常環境との違いは，シミュレーション開始から 10,000 step 後に必ず全ての選択肢の真の報酬生起確率が一様分布から独立に再設定される事である．選択肢の変化のタイミングが同期しているため，同期的非定常環境と呼ぶ．各エージェントは選択肢の変化を直接的に観測する事は出来ず，また，そのタイミングを学習する能力も持たない．シミュレーション 1,000 回のあいだ毎回，全ての選択肢の真の報酬生起確率は設定しなおされる．比較するアルゴリズムには，LSX, LSX_{opt}, $UCB1\text{-}tuned$(全て $\gamma = 1.0$ と 0.999 の両方について，$\gamma = 0.999$ を用いた場合はモデ

ル名 γ と表記する) とメタバンディットアルゴリズムを用いた．LSK は定常バンディット問題において著しく成績が悪く．また LSX_{cc} は定常バンディット問題のためのアルゴリズムであるため非定常バンディット問題シミュレーションからは除外した．

5.4.1 シミュレーション 2A の結果

以下に定常環境におけるシミュレーション 2A の結果を示す．図 5.4 は正解率の推移であり，図 5.5 は後悔の度合いの推移であり，図 5.6 は入替率の推移である．

図の見方は前述の通りだが，図 5.4 において，ほぼ全てのアルゴリズムに対して 10,000 step 毎に正解率の減少が見られる事がわかる．これは同期的に環境が変化するタイミングと一致しており，各アルゴリズムはその変化を選択肢の選択と報酬として現れる結果の変動のみで感知して対応しているように見える．最も変化によく対応しているのは 忘却率 $\gamma = 0.999$ によって過去の情報を圧縮して更新している場合の オンライン LSX アルゴリズムと最適基準 LSX アルゴリズムだった．この二つのアルゴリズムのみ，環境の変化前と同じ高い水準まで正解率を回復する事が出来ている．環境の変化前と変化後の正解率がほぼ変わらない点であれば忘却率 $\gamma = 0.999$ の $UCB1\text{-}tuned$ も同様であるが，前述した二つと比べてかなり成績が低い．非定常バンディット問題のためのアルゴリズムであるメタバンディット ($Meta\ UCB1\text{-}tuned$) も 60% 程まで正解率を回復できているものの，LSX には劣る．情報の圧縮をせずとも $UCB1\text{-}tuned$，オンライン LSX，最適基準 LSX は正解率を回復している傾向がある．特に最適基準 LSX は非圧縮としては優れた成績を有しているが，徐々にその回復量は落ち込んでいる事がわかる．これは図 5.6 に示す入替率にも表れており，選択肢の入替率から見られる環境の変化によって発生する探索行動の割合が，前述した 3 つアルゴリズムにおいて環境変化の度に大きくなっている．その中で圧縮したオンライン LSX と最適基準 LSX のみは度重なる変化を迎えても，一定で非常に少ない入れ替え率で，図 5.1 に表された正解率の回復を達成しているのがわかる．これらの傾向は図 5.5 にも表れており，やはり $\gamma = 0.999$ によって過去の情報を圧縮して更新している場合の オンライン LSX アルゴリズムと最適基準 LSX アルゴリズムが最も低い水準を保っている．表 5.2 に示す通り，この二つのアルゴリズムのみが後悔の度合いを 1,000 以下に抑えられ，オンラインで学習するアルゴリズムとしては実質的にオンライン LSX アルゴリズムのみである．

忘却率 γ を用いる事で，非定常環境に対する対応力が向上するのは，ある側面で言えば当然の事だと考えられる．前述した通り，忘却率を用いた更新は式 4.73 に示される新しく獲得した情報に対する重みの保証を有しているため，古い情報の重みに引っ張られる割合が相対的に減少する．故に同期的な環境の変化への対応力は向上するのだが，それだけでは非定常環境に対応しきれない．実際に本シミュレーションにおける情報を圧縮した $UCB1\text{-}tuned$ の結果は著しく低下している．シミュレーション 1 でも述べた通り，忘却率 γ による情報の圧縮は，各アルゴリズムが定義する価値を不安定にする．特に統計的な強い制約を持つ UCB1-tuned では圧縮する事を想定していないために制約を満たす事が出来ず，本来の想定通りに機能していない．過去の情報の圧縮は長期的な情報の積み上げによって得られる保証を満たす事ができなくなるため，そのような保証とは異なる方略を有するアルゴリズムでなければ扱う事が出来ない．シミュレーション 1 に示したオンライン LSX の成績から，それが長期的な保証を優先するアルゴリズムでないことがわかる．また定常環境においても忘却率 γ を用いる事で成績が減少しない事を示しているため，忘却率による情報の圧縮がオンライン LSX においては非定常環境に対応するためのアドホックな調整ではない事がわかる．また，忘却率による過去の情報の圧縮とそれに付随する新しい情報に対する重み (情報全体に対する影響度合い) の最低保証は非常に直

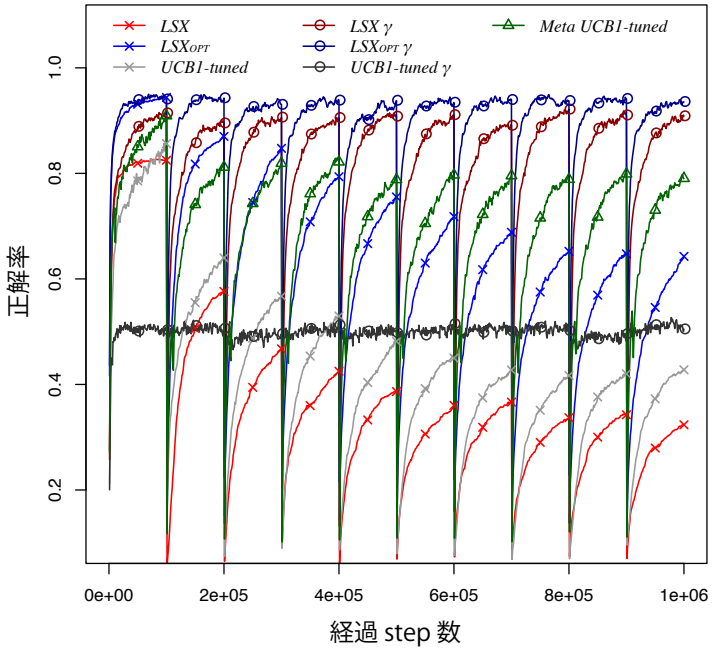

図 5.4: 非定常・同期 20 本腕バンディット問題：正解率

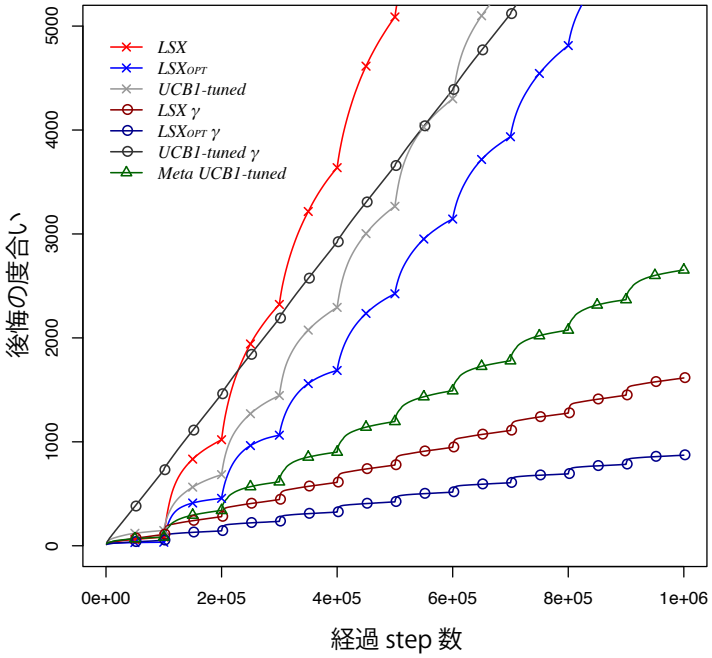

図 5.5: 非定常・同期 20 本腕バンディット問題：後悔の度合い

感的な情報保存の形式であると考えられる．このような直感的な調整パラメータに対応できるのも，*LSX* が認知的で直感的な方略を有するアルゴリズムである事に起因すると考えられる．

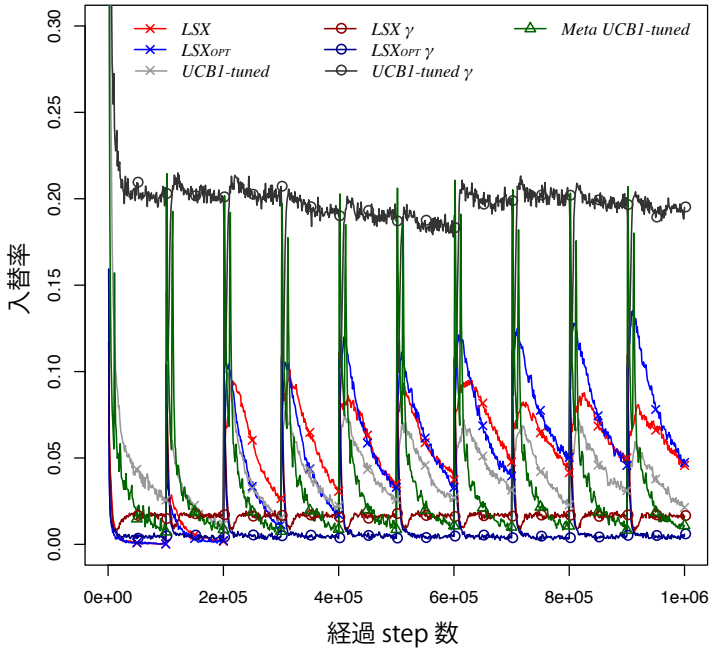

図 5.6: 非定常・同期 20 本腕バンディット問題：入替率

表 5.2: 非定常・同期 20 本腕バンディット問題：99,900 ～ 100,000 step の指標平均値

	正解率	後悔の度合い
LSX	0.324	12701.4
LSX γ	0.908	1614.5
LSX_{opt}	0.643	6621.7
LSX_{opt} γ	0.936	871.7
UCB1-tuned	0.428	8853.4
UCB1-tuned γ	0.505	7308.1
Meta UCB1-tuned	0.790	2653.4

5.5 シミュレーション 2B：非定常-非同期

前述した共通設定の通り，本シミュレーションに選択を行うエージェントは，選択アルゴリズムに従い選択を総 step 数 として 100,000 steps 行い，そのシミュレーション 1,000 回分を平均して各種指標を算出した．選択肢は 20 通りあり，全ての選択肢の真の報酬生起確率は，シミュレーション開始時に一様分布から独立に決定される．前述したシミュレーション 1 の定常環境やシミュレーション 2A の同期的非定常環境との違いは，シミュレーション開始から，毎 step 毎に 10,000 分 1 の確率で各選択肢それぞれの真の報酬生起確率が一様分布から独立に再設定される事である．同期的非定常環境と異なり，選択肢の変化のタイミングが同期していないため，非同期的非定常環境と呼ぶ．各エージェントは選択肢の変化を直接的に観測する事は出来ず，また，そのタイミングを学習する能力も持たない．シミュレーション 1,000 回のあい

表 5.3: 非定常・非同期 20 本腕バンディット問題：99,900 ～ 100,000 step の指標平均値

	正解率	後悔の度合い
LSX	0.261	12225.6
LSX γ	0.763	1833.9
LSX_{opt}	0.540	6375.0
LSX_{opt} γ	0.888	1235.6
*UCB*1-*tuned*	0.304	8509.1
*UCB*1-*tuned* γ	0.512	7329.3
*Meta UCB*1-*tuned*	0.573	3200.5

だ毎回，全ての選択肢の真の報酬生起確率は設定しなおされる．比較に用いたアルゴリズムはシミュレーション 2A(非定常-同期) と同様である．

5.5.1　シミュレーション 2B 結果

以下に定常環境におけるシミュレーション 2A の結果を示す．図 5.7 は正解率の推移であり，図 5.8 は後悔の度合いの推移であり，図 5.9 は入替率の推移である．

図の見方は前述の通りだが，図 5.7 において，基本的に全ての選択肢の正解率が徐々に減衰していくことがわかる．シミュレーション 2A の同期的定常環境は，環境の変化が必ず全ての選択肢に発生し，かつ周期的に行われる点に特徴が合った．これは隕石や大恐慌等による広範囲の環境変化を想定するならばあり得る環境ではあるが，基本的にはそれぞれの選択肢の変化が他の変化と同期的に行われるとは限らない．そのため非定常な環境を想定するなら非同期的で確率的だと考えるのが自然であるが，この課題の特徴はそのような自然さだけで表されるわけではない．選択肢の変化に対する課題の難しさは，例えば充分に情報が集まり，選択する選択肢がほぼ一つに決まった場合 (前述した選択収束状態) において，変化が同期的か非同期的かで大きく変わる．同期的環境なら，選択収束状態で現在選択し続けている選択肢に対する観測結果の変化から，それを感知して他の選択肢を再探索することが可能である．しかしながら，変化が非同期である場合，現在エージェントが選択していない他の報酬生起確率が低かった選択肢が，変化によって高くなったとしても，その変化を検出できないし，そのために探索を行おうともしない．シミュレーション 2A の同期的非定常環境で高い成績を有していた忘却率 $\gamma = 0.999$ によって過去の情報を圧縮して更新している場合の オンライン *LSX* アルゴリズムと最適基準 *LSX* アルゴリズムは，非同期的非定常環境でもよく振る舞った．正解率は高い水準を示し，後悔の度合いでも表 5.3 に示される通り，上記の二つのアルゴリズムのみ 2,000 を下回っている．非定常バンディット問題のためのアルゴリズムであるメタバンディット (*Meta UCB*1-*tuned*) も通常の．*UCB*1-*tuned* と比較すると後悔を著しく低下させているものの，*LSX* には劣る．これはメタバンディットが対応できるのは，最も高いと観測される選択肢の価値が減少する場合の非定常さであるからと考えられる．また図 5.9 に示される入替率の推移おいても，シミュレーション 1 の定常環境における入替率に比べて環境に応じてより高く調整している事がわかる．前述した通り，最適基準 *LSX* は未知な環境に対して何の事前知識の無い状態で扱えるアルゴリズムではないため，動的な学習アルゴリズムとしては *LSX* のみが良い結果を示した．

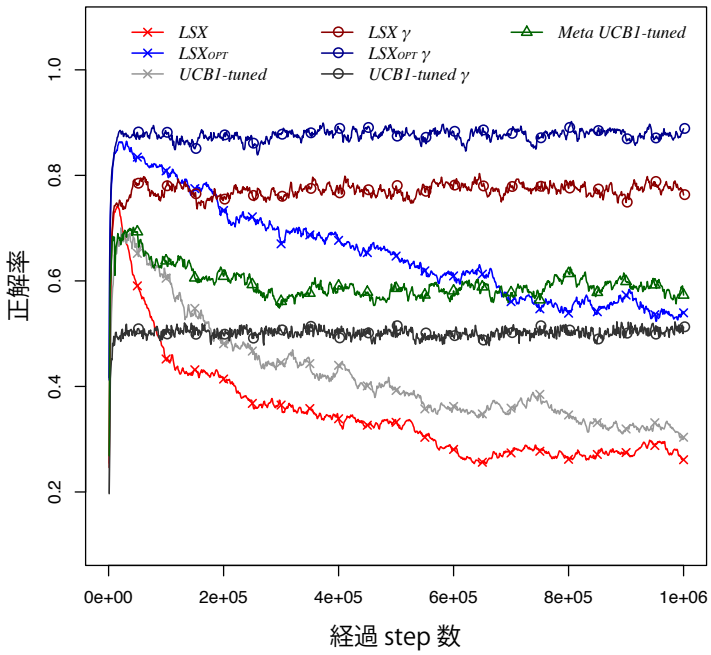

図 5.7: 非定常・非同期 20 本腕バンディット問題：正解率

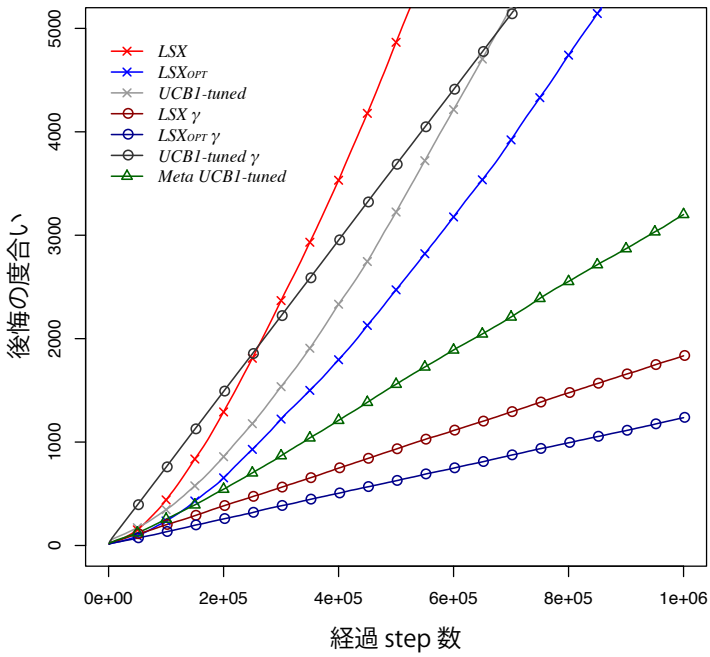

図 5.8: 非定常・非同期 20 本腕バンディット問題：後悔の度合い

5.6 考察

我々が考案した *LSX* は強化学習課題における主観的な価値関数としての *LS* の欠点を補うよう拡張した数理モデルである．同時に基準値の漸進的な学習等と組み合わせる事により，意思

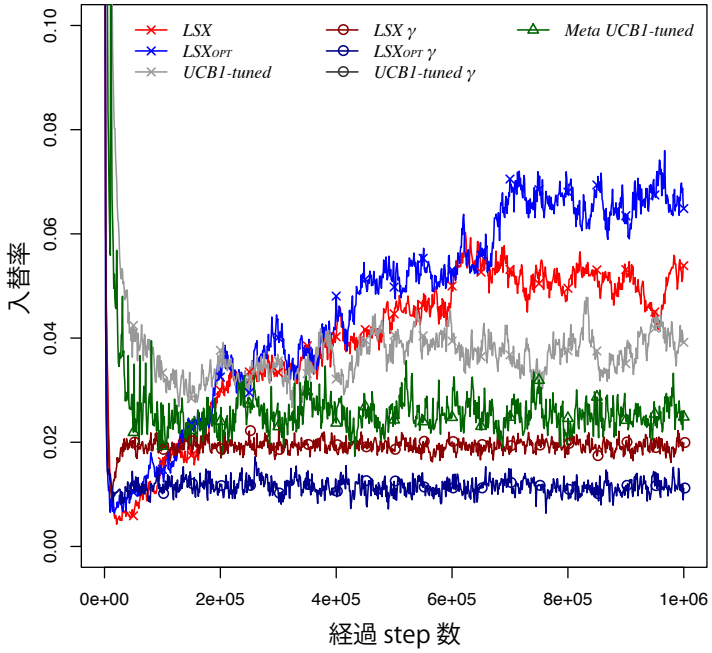

図 5.9: 非定常・非同期 20 本腕バンディット問題：入替率

決定課題，強化学習課題に対する実用的なアルゴリズムとしても機能する．

本研究では [Oaksford 94]，[Tenenbaum 11] 等から，人間の認知的特性が事前知識を持たないエージェントが未知の環境に対して適応する時に有効であるという前提を有している．本研究ではそれらに属する特定の認知的特性が有効である事をモデルを用いてシミュレーション上で実証した．特に同期的，あるいは非同期的にその因子が変化する非定常な環境において，単純なシミュレーションから有用性を示す事が出来た．そのような非定常環境において，情報を圧縮して新しい情報の反映度合いを最低限保ち続けるという情報の保存形式が重要になる．*LSX* は情報の蓄積による定常環境において正しい選択を行う事を目的としていないために，そのような情報の圧縮と相性がよく，また圧縮しなくとも基準値が最適であるならば非定常環境によく対応することが本研究のシミュレーションに表れている．それにより，*LSX* がメタバンディットアルゴリズムでは対応できなかったような最も高い選択肢以外が変化していく環境にも対応できるという新たな知見が得られた．

第6章　強化学習と満足化

以上の性質をふまえ，*LS* を方策関数として強化学習に応用する方法を提示する．強化学習とは環境と学習エージェントの相互作用から最適な方策を学習する機械学習の一種である．強化学習においてエージェントは自身と環境の状態 s を把握し，その状態 s において可能な行動群 A の中から実際に取る行動 a を意思決定方策に基づき選択し，次の状態 s' に推移．報酬 r を得る．これらの情報から状態行動対 (s,a) の価値，そして方策 π を学習する．

強化学習と教師あり学習，教師なし学習の最大の相違点は，エージェント自身が主体的に環境から情報を探索しなければならない事にある．報酬を得る事のみを優先すれば探索は進まず，かといって情報探索に時間を費やせば最終的に得られる報酬は少なくなってしまう．他にも強化学習の抱える問題は連続状態，連続行動に対する価値関数の近似等，様々存在する，しかし収穫と探索のジレンマを抱えている以上，少なくともその問題においては LS あるいは LSX を強化学習に用いる事は有用であると思われる．

6.1　TD 学習

強化学習の代表的学習手法である TD 学習はマルコフ過程を前提とし，Q 値と呼ばれる状態行動対 (s,a) に対応する行動価値関数 $Q(s,a)$ を以下の Bellman 方程式に従い学習する．

$$Q(s_t,a_t) \quad \leftarrow \quad Q(s_t,a_t) + \alpha[r + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_{up})] \tag{6.1}$$

特に以下の式 6.1 を *TD* 差分と呼び，現在の価値と得られた情報の差異を意味している．この *TD* 差分を一つ前の状態にバックアップする事で，オンラインに学習する事が出来る．

$$[r + \gamma Q(s_{t+1},a_{up}) - Q(s_t,a_t)] \tag{6.2}$$

ここで，α は得られた情報をどれほど学習するかを決める学習率 $(0.0 \leq \gamma \leq 1.0)$，$\gamma$ は未来に得られる価値をどれだけ減衰させるかという割引率である $(0.0 \leq \gamma \leq 1.0)$．割引率は低い程即時報酬，即ちすぐに手に入る報酬に価値を見いだすようになり，高ければ将来手に入る報酬を高く見積もるようになる．

ここでバックアップに使う行動 a_{up} の選び方で TD 学習の名称が変わる．実際の行動を選択する際に意思決定方策を反映させるのみでなく，バックアップする行動 a_up にも意思決定方策に従って選択した次に取る行動 a_{t+1} を反映させる強化学習アルゴリズムを *Sarsa* と呼ぶ．*Sarsa* はバックアップ時にも意思決定方策を活用する事から方策オン型の TD 学習．それに対して方策オフ型 TD 学習では Q 学習が知られている．Q 学習と *Sarsa* 違いはバックアップに使う行動には常に価値価数が最大の物を選択する (*greedy*) 事のみである．

6.2　強化学習とトレードオフ

基本的な強化学習アルゴリズムの一種である TD 学習 (Q 学習，Sarsa 等) は，獲得した報酬を Q 値と呼ばれる任意の状態においてある行動を取る事 (状態行動対) に対する価値として格納す

る．TD 学習では Q 値と選択方策を参照して次に取るべき行動を選択し，それが現時点での Q 値に照らして合理的であった場合は利益追求行動，そうでない場合を探索行動と定義される．最も目的に沿った行動系列を発見するためには，なるべく多くの報酬に関する情報を探索する必要がある．しかしながら探索する程に報酬を直接的に得る機会は失われて行く．このように探索と利益追求は両立できないため，報酬獲得の速さ (Speed) と高い報酬を得るための Q 値の正確さ (Accuracy) にはトレードオフの関係がある (Speed and Accuracy Trade-off)[Wickelgren 77]．実際に選択される行動は方策アルゴリズムによって決定される．例えばシンプルな方策アルゴリズムである ϵ-greedy では，確率 ϵ でランダムな行動を行い，残りの $(1-\epsilon)$ 確率で利益追求行動を行うという乱数を用いた選択を行う．このように，トレードオフ に対処して良い探索と利益追求のバランシングを行えるか否かは行動選択の方策アルゴリズムが担う問題となる．そのシンプルさから ϵ-greedy は扱い易いが，ϵ が一定である場合はいつまでも確率的に探索が発生する上に，その減衰はパラメータの経験や解析による設計が学習課題毎に個別に必要となる．より効率的な探索を目的とした Bolzman 分布を応用した softmax 方策等も存在するが，ϵ が減衰する ϵ-greedy より更に複雑なパラメータ設計が必要となる．そこで我々は人間の柔軟な意思決定に習う事で，柔軟に探索と利益追求を配分し，かつ扱い易い方策アルゴリズムを作れないかと考えた．

6.3　強化学習と満足化

前述の K 本腕バンディット問題に対するシミュレーションと考察から，LSX が定常・非定常の不確実さを持つ環境での選択で良い振る舞いをする事がわかった．以降ではより発展的な学習課題である強化学習への LSX の応用を行う．事前知識が無い状態で目的を達成するためには試行錯誤して情報を収集する事が重要となる．環境に対する試行錯誤と得られる報酬から合目的的行動を獲得する枠組みである強化学習では，情報収集のための行動選択を “探索” と定義している．他方，得られた情報から最も目的達成に近づけると見積もられる行動を選択する事を “利益追求” と呼ぶ [Sutton 00]．合目的的行動の獲得に際する大きな問題の一つが，この探索行動と利益追求行動のバランシングである．あらゆる行動パターンを全て網羅しつつ，確率的な不確実性まで考慮して環境を探索し尽くせるならともかく，大抵の場合，学習エージェントやタスクには時間や行動コストに対する量的な制限があるため，無限に近い時間を探索に割り当てる事は出来ない．その学習課題の性質にもよるが，現実的な環境であればある程，学習エージェントにはなるべく短い時間で，なるべく良い行動を獲得する事が求められる．そのため，探索行動の配分は強化学習においての根幹に関わる問題であると言える．

ゆえに強化学習アルゴリズムに価値関数として LSX を導入する事は有用であると考えられる．しかしながら，LSX アルゴリズムは報酬の生起確率のみに対応しており，実数値域の報酬や価値には対応していなかった．そこで本研究では，より広い強化学習課題に応用する際の LSX の問題点を考察し，それを改善した Real scaLize Loosely Symmetric model (以下 RLLS) を考案した．また複雑な強化学習課題における RLLS アルゴリズムの振る舞いを通して LS 系モデルの根幹的な性質 (満足化) がどのような効果を生んでいるか考察した．

6.4　Real scaLize Loosely Symmetric model

前述では LSX は客観的にえられた割合を信頼性や基準値に応じて値を歪めさせる価値関数として定義してきた．あくまで扱うのは割合であり，互いに独立な選択肢 ($a_j \in A$) と，その選択

肢を試行した際に観測された目的事象 (報酬，$e_k \in \{e, \bar{e}\}$) の割合変数 $S(a_j \cap e_k)$ 等の表記を元にモデルの記述を行ってきた．しかし強化学習ではある選択肢から得られる報酬獲得割合 $S(e|a_i)$ に相当する行動価値関数 (Q 値) と対応づける必要がある．Q 値はある方策 π の時に状態 s_i で行動 a_j がその後えられる収益 (報酬 r の累積) の期待値を意味するため，以降は期待値 $\overline{X_{a_i}}$ とその試行回数 n_{a_i} によって表記する．これは強化学習への対応を考慮するのみでなく，より一般的な表記とすることで広い範囲での応用の示唆が得られる．期待値 $\overline{X_{a_i}}$ とその観測量 (試行の度合い)n_{a_i} により表記し直すと，価値関数 LSX の評価値は以下の式によって定義される．

$$a_{MT} = \arg\max_{a_k}(n_{a_k}) \tag{6.3}$$

$$a_{LT} = \arg\min_{a_k}(n_{a_k}) \tag{6.4}$$

$$V_e = \frac{n_{a_{MT}}\overline{X_{a_{MT}}} n_{a_{LT}}\overline{X_{a_{LT}}}}{n_{a_{MT}}\overline{X_{a_{MT}}} + n_{a_{LT}}\overline{X_{a_{LT}}}} \tag{6.5}$$

$$V_{\bar{e}} = \frac{n_{a_{MT}}(1-\overline{X_{a_{MT}}}) n_{a_{LT}}(1-\overline{X_{a_{LT}}})}{n_{a_{MT}}(1-\overline{X_{a_{MT}}}) + n_{a_{LT}}(1-\overline{X_{a_{LT}}})} \tag{6.6}$$

$$n_V = V_e + V_{\bar{e}} \tag{6.7}$$

$$X_V = \frac{V_e}{n_V} \tag{6.8}$$

$$\omega_{n_i} = \frac{n_i}{n_i + n_V} \tag{6.9}$$

$$LSX(e; a_i) = \omega_{n_i}\overline{X_{a_i}} + (1-\omega_{n_i})(2R - X_V) \tag{6.10}$$

$$a_{Select} = \arg\max_{a_k}\left(LSX(e; a_k)\right) \tag{6.11}$$

行動選択は LSX の評価値 (式 6.10) において最大となる行動選択肢が選ばれる (式 6.11) ため，評価値がまったく等しい場合以外には乱数を用いず，エージェント側としては決定論的に行動が選択される．また前述した通り，行動 a_i の全試行回数に対する試行割合 T_{a_i} が 1.0 に近づいた時，即ち試行全体において既にほぼ探索的行動をしておらず，ある行動選択肢に執着している際には LSX の評価値は客観的な評価と一致する (鮮明化，式 6.14)．逆に試行割合 T_{a_i} が 0.0 に近づいた時，即ち相対的にほとんど選択されていない場合は最も曖昧な評価値である基準値 R に収束する (背景化，6.15).

$$T_{a_i} = \frac{n_{a_i}}{\sum_{a_k \in A} n_{a_k}} \tag{6.12}$$

$$0.5 < \omega_{n_i} \leq 1.0 \tag{6.13}$$

$$\lim_{T_{aH}\to 1.0} LSX(e; a_{MT}) = X_{a_{MT}} \tag{6.14}$$

$$\lim_{T_{aL}\to 0.0} LSX(e; a_{LT}) = R \tag{6.15}$$

6.4.1　価値関数としての LSX の問題

強化学習，正確には実数値の値域にそのまま拡張することを困難にする大きな問題が，LSX には二つ存在する．第一には良い基準値 R を如何にして獲得するかであり，第二の問題は客観的な価値である標本平均 $\overline{X}$ を Q 値のような実数値の範囲にすると背景化等の性質が失われること等が挙げられる．本研究ではまず第二の問題について議論する．LSX の評価値は経験的に得られた客観的価値 $\overline{X}$ と基準値 R，仮想的な価値 X_V との重み付け平均によって算出される [甲野 14]．性質が失われる理由は，この際に用いられる重み ω_{n_i} が，価値 $\overline{X}$ が正負に股がった実数値を扱う際に重みとしての性質を破綻させるためである．具体的には V_e と $V_{\bar{e}}$ の分母に価値 $\overline{X}$ が含まれる事が原因である．

$$n'_V = \frac{n_{a_{MT}} n_{a_{LT}}}{n_{a_{MT}} + n_{a_{LT}}} \tag{6.16}$$

$$w_{V_e} = \frac{n_{a_{MT}} \overline{X_{a_{MT}}}}{n_{a_{MT}} \overline{X_{a_{MT}}} + n_{a_{LT}} \overline{X_{a_{LT}}}} \tag{6.17}$$

$$w_{V_{\bar{e}}} = \frac{n_{a_{MT}} (1 - \overline{X_{a_{MT}}})}{n_{a_{MT}} (1 - \overline{X_{a_{MT}}}) + n_{a_{LT}} (1 - \overline{X_{a_{LT}}})} \tag{6.18}$$

$$V_e = n_V \left(w_{V_e} \overline{X_{a_{LT}}} + (1 - w_{V_e}) \overline{X_{a_{MT}}} \right) \tag{6.19}$$

$$V_{\bar{e}} = n_V \left(w_{V_{\bar{e}}} (1 - \overline{X_{a_{LT}}}) + (1 - w_{V_{\bar{e}}}) (1 - \overline{X_{a_{MT}}}) \right) \tag{6.20}$$

$$n'_V \neq V_e + V_{\bar{e}} = n_V \tag{6.21}$$

$$w_{V_e} \neq w_{V_{\bar{e}}} \tag{6.22}$$

そこで我々は仮想的な試行回数 n_V から価値 $\overline{X}$ を排除した n'_V(式 6.16) と置き換えて式変形を行い，n_V と n'_V が等しくならない事 (式 6.21) の原因が式 6.22 にあると推定した．

6.4.2　LSX から RLLS への拡張

我々は前述の推定に基づき，w_{V_e}，$w_{V_{\bar{e}}}$ を等しく w'_{V_e} に差し替える事で重み ω'_{n_i} から価値 $\overline{X}$ を排除した．これにより価値 $\overline{X}$ が正負にまたがった実数値の範囲を取った場合にも鮮明化 (式 6.33) と背景化 (式 6.34) の性質を保つことが出来るようになった．我々はこの評価式を Real scaLize Loosely Symmetric model (RLLS) と名付けた [甲野 15]．

$$w'_V = \frac{n_{a_{MT}}}{n_{a_{MT}} + n_{a_{LT}}} \tag{6.23}$$

$$V'_e = n_V \left(w'_V \overline{X_{a_{LT}}} + (1 - w'_V) \overline{X_{a_{MT}}} \right) \tag{6.24}$$

$$V'_{\bar{e}} = n_V \left(w'_V (1 - \overline{X_{a_{LT}}}) + (1 - w'_V) (1 - \overline{X_{a_{MT}}}) \right) \tag{6.25}$$

$$n'_V = V'_e + V'_{\bar{e}} = \frac{n_{a_{MT}} n_{a_{LT}}}{n_{a_{MT}} + n_{a_{LT}}} \tag{6.26}$$

$$X'_V = \frac{V'_e}{n'_V} \tag{6.27}$$

$$\omega'_{n_i} = \frac{n_i}{n_i + n'_V} \tag{6.28}$$

$$RLLS(e;a_i) = \frac{n_{a_i}\overline{X_{a_i}} + n'_V 2R - V'_e}{n_{a_i} + n'_V} \tag{6.29}$$

$$= \omega'_{n_i}\overline{X_{a_i}} + (1-\omega'_{n_i})(2R - X'_V) \tag{6.30}$$

$$RLLS(e;a_i) = \omega'_{n_i}\overline{X_{a_i}} + (1-\omega'_{n_i})(2R - X'_V) \tag{6.31}$$

$$0.5 < \omega'_{n_i} \leq 1.0 \tag{6.32}$$

$$\lim_{T_{aH}\to 1.0} RLLS(e;a_{MT}) = X_{a_{MT}} \tag{6.33}$$

$$\lim_{T_{aL}\to 0.0} RLLS(e;a_{LT}) = R \tag{6.34}$$

6.5 強化学習における RLLS 方策

RLLS により価値 $\overline{X}$ を Q 値のような実数値に置き換える事は可能となった．しかしながら，TD 学習には観測量 n に相当する概念が存在しないため，その導入が必要となる．そこで我々はある状態 s_i において a_j を試行した強さとして信頼度変数 τ 値 (変数 $\tau(s_i,a_j)$) という量を定義した (表 6.1).

表 6.1: 状態 s_i における Q 値と τ 値

	Q	τ
a_1	$Q(s_i,a_1)$	$\tau(s_i,a_1)$
a_2	$Q(s_i,a_2)$	$\tau(s_i,a_2)$
$\vdots$	$\vdots$	$\vdots$
a_n	$Q(s_i,a_n)$	$\tau(s_i,a_n)$

Q 値の更新は従来通り方策 on 型 TD 学習を用いるか方策 off 型 TD 学習をベースとするかによって更新法が異なる．強化学習アルゴリズムにおける RLLS は式 6.31 において価値 $\overline{X}$ を Q 値に，観測量 n を τ 値に置き換える事で定義できる (式 6.39)．ここで基準値 R_i は状態 s_i 毎に個別の値を持つ.

$$a_{MT} = \arg\max_{a_k}\left(\tau(s_i,a_k)\right), a_{LT} = \arg\min_{a_k}\left(\tau(s_i,a_k)\right) \tag{6.35}$$

$$Q_u = \frac{\tau(s_i,a_{MT})Q(s_i,a_{LT}) + \tau(s_i,a_{LT})Q(s_i,a_{MT})}{\tau(s_i,a_{MT}) + \tau(s_i,a_{LT})} \tag{6.36}$$

$$\tau_u = \frac{\tau(s_i,a_{MT})\tau(s_i,a_{LT})}{\tau(s_i,a_{MT}) + \tau(s_i,a_{LT})} \tag{6.37}$$

$$\omega_{ij} = \frac{\tau(s_i,a_j)}{\tau(s_i,a_j) + \tau_u} \tag{6.38}$$

$$RLLS(s_i,a_j) = \omega_{ij}Q(s_i,a_j) + (1-\omega_{ij})(2R_i - Q_u) \tag{6.39}$$

強化学習における RLLS 方策も LSX と同様に RLLS 価値関数が最も高い行動を選択する (式 6.40)．この選択は前述の通り鮮明化と背景化の性質も有しているため，乱数を用いずに満足化方策として機能する．

$$a_{Select} = \arg\max_{a_k} \left(RLLS(s_i, a_k) \right) \tag{6.40}$$

$$T_{ij} = \frac{\tau(s_i, a_j)}{\sum_{a_k \in A_{s_i}} \tau(s_i, a_k)} \tag{6.41}$$

$$0.5 < \quad \omega_{ij} \quad \leq 1.0 \tag{6.42}$$

$$\lim_{T_{iH} \to 1.0} RLLS(s_i, a_{MT}) \quad = \quad Q(s_i, a_{MT}) \tag{6.43}$$

$$\lim_{T_{iL} \to 0.0} RLLS(s_i, a_{LT}) \quad = \quad R \tag{6.44}$$

6.5.1　τ 値の定義と更新手法

信頼度変数 τ 値は状態価値関数 Q 値の信頼度を意味する変数である．Q 値はある方策 π の下で，状態 s_i の時に行動 a_j を取った場合，その後得られる収益の近似値を意味する．そのため信頼度変数 τ 値は方策 π の下で，状態 s_i の時に行動 a_j を経験した回数等に相当すべきである．しかし，一般的な TD 学習において方策 π は学習中に比較対象となる行動間の Q 値の大小関係が変化する毎に変化する．つまり，τ 値が表すべき方策 π に関する情報の観測度合い，蓄積度合いを直接評価することはできない，

そこで本研究では RLLS において τ 値の相対的な比率が重要なことに因み，現時点での π が生成するであろう系列から τ 値を定義し，更新する事とする．

具体的には τ 値を現状態 s_i での行動 a_j を行った回数を $\tau_{current}(s_i, a_j)$(現行動信頼度変数) と，その後の状態行動系列を信頼度変数を $\tau_{post}(s_i, a_j)$(未来信頼度変数) の和によって定義する．

$$\tau(s_i, a_j) = \tau_{current}(s_i, a_j) + \tau_{post}(s_i, a_j) \tag{6.45}$$

$$\tau_{current}(s_t, a_t) \quad = \quad \tau_{current}(s_t, a_t) + 1 \tag{6.46}$$

$$\tau_{post}(s_t, a_t) \quad = \quad \tau_{post}(s_t, a_t) + \alpha_\tau \left(\gamma_\tau \tau(s_{t+1}, a_{up}) - \tau_{post}(s_t, a_t) \right) \tag{6.47}$$

$\tau_{current}$ と τ_{post} の更新はそれぞれ，式 6.46，6.47 によって行われる．更新に用いる a_{up} は方策 on 型の TD 学習アルゴリズムであれば実際に方策に従い選択された行動 a_{t+1} を用い，方策 off 型であれば $\arg\max_{a_k} Q(s_{t+1}, a_k)$ となる．未来信頼度変数 τ_{post} は信頼度学習率 $\alpha_\tau = 1.0$，未来系列信頼度割引率 $\gamma_\tau = 1.0$ の時，その後の系列の試行回数の累積となる．信頼度学習率 $\alpha_\tau = 1.0$ である場合，その後の系列に変化があれば，累積を計算するために参照する状態行動対の組み替えが起こる．即ち信頼度学習率 α_τ は Q 値の学習率 α_Q と同様，組み変わった状態行動対の組み替えをどの程度抑制するかを意味するパラメータである．そのため基本的には Q 値の学習率 α_Q と信頼度学習率 α_τ は等しくなるべきだと考えられる．また，未来系列信頼度割引率 $\gamma_\tau (0 \leq \gamma_\tau \leq 1)$ はその後の状態行動対の訪問回数を割り引くために扱う．即ち未来系列信頼度割引率 γ_τ は選択肢 a_j を選んだ後の試行回数をどの程度考慮するかを意味している．

6.6 強化学習における基準値とその獲得

K 本腕バンディット問題と同様 (式 4.76) に，強化学習における RLLS にも適した基準値の条件が存在する (式 6.48)．ただし，基準値 R は状態 s_i 毎に存在し，適した基準値 R も状態 s_i 毎に存在する．

$$Q^*_{Second}(s_i) < \quad R(s_i) \quad < Q^*_{First}(s_i) \tag{6.48}$$

ここで Q^* は最適な方策 π^* で獲得できる最適価値関数であり，$Q^*_{First}(s_i)$ は状態 s_i において一番高い最適価値関数 Q^* の値であり，$Q^*_{Second}(s_i)$ は同様に二番目に高い値である．TD 学習の最適価値関数 Q^* は K 本腕バンディット問題における真の報酬獲得確率に相当するが，多くの場合，課題設計者が任意に決定できる値ではなく，また，複雑な環境であると計算も困難になる．

状態 s_i 毎の基準値 R の動的な学習も K 本腕バンディット問題と同様に可能である．しかし K 本腕バンディット問題で用いていた漸進的基準値更新法 (式 4.74) を Q 値に適用する (式 6.49) のみでは強化学習課題で良い成績は得られない．

$$R(s_i) \leftarrow R(s_i) + \alpha_R(Q(s_i, a_{Select}) - R(s_i)) \tag{6.49}$$

K 本腕バンディット問題で漸進的基準値更新法が有効だったのは，期待値 (報酬獲得割合 $S(e|a_j)$) が選択肢 a_j に対して十分試行すれば真の報酬獲得割合 P_{a_j} に収束する事に起因する．しかし強化学習の期待値に相当する Q 値は環境の不確実性のみならず，その後の系列をどう振る舞うか，端的には方策 π に依存するため，試行回数を増やしたところで最適価値関数 Q^* には収束しない．故に，単純な漸進的価値更新法では良い基準値 R を得る事はできない．

ここで満足化の基本的な性質に立ち返る．満足化において探索と利益追求は，基準値 R を上回る価値 (Q 値) を持った行動がなければ探索，上回る価値があれば利益追求という条件で使い分けられていた．つまり $Q^*_{First}(s_i)$ を発見するためは，現場得られている Q 値より基準値が高くあり続ければ良い事になる．方策 π が変わらなくても，次状態の Q 値を参照するバックアップ更新の性質から，Q 値の更新速度は非常に遅い．それに対して，最終的に Q 値が近似すべき L step 収益を直接観測して更新に用いる強化学習アルゴリズムである Q-timer は TD 学習のバックアップ更新よりも (正確性が損なわれる代わりに) 更新速度が速い [太田 14]．そのため，Q-timer アルゴリズムの Q 値更新法に習った R の基準値更新を考案した．

6.6.1 R-timer 基準値更新

Q-timer は状態行動対 (s, a) に訪問した後 L step の間に得られた報酬の累積，L step 収益 (R_L) を用いて行動価値関数 $Q(s, a)$ を更新する手法である．本来終端状態まで観測すべき収益の観測を L step で切り上げる代わりに TD 学習のバックアップ更新よりも 現方策 $\pi_{current}$ で得られる収益を即座に Q 値に反映する事ができる．しかし，L step 収益であるため，TD 学習のバックアップ更新よりも遅延報酬を見逃しやすい．L step を十分長くすれば正しく収益を観測できるが，長過ぎれば極端に Q 値の更新頻度が下がってしまう．また，問題によって正しい L step の長さが異なるため，汎用性にかける．

$$R_L(s_t) = r_t + r_{t+1} + ... + r_{t+L} \tag{6.50}$$

しかし，Q 値への更新では正確性が重視されるが，基準値 R への更新は前述の問題から速さが重視される．特に満足化の性質から特に上昇するようなに対して速くなるべきである．そのため，L step 収益を用いて R を式 6.51 によって更新する手法を考案した．

$$R(s_t) = \begin{cases} R(s_t) & (R(s_t) \geq R_L(s_t)) \\ R_L(s_t) & (R(s_t) < R_L(s_t)) \end{cases} \tag{6.51}$$

ただし L step 収益 R_L は不確実であるため，Q 値を用いた漸進的基準値更新も併用する (式 6.49)．ただし，飽くまでもこの漸進的基準値更新は R_L の不確実さを安定させる事を目的としているため，学習率 α_R は非常に低い値とする．これは R_L が偶然高すぎる値になってしまった場合に対応するため，基準値という目標値を現在得られている Q 値に馴らしている事を意味する．そのためこの漸進的な更新を馴化と呼び，学習率 α_R を馴化率と呼ぶ．また L step 収益による更新と馴化を組み合わせたこの基準値の更新法を，R-timer 基準更新と名付けた．

6.7　強化学習シミュレーション

RLLS 方策を用いた学習アルゴリズムの性質を検証するため，本研究では複雑なダイナミクスを有する運動課題である大車輪強化学習課題を用いる．大車輪課題を用いた理由は，RLLS とは異なる価値関数 LS の強化学習の応用案 (LS-Q アルゴリズム [浦上 13]，LS-VR-Q アルゴリズム [高橋 13]) で検証されている課題である事も考慮している．LS-Q アルゴリズムやその発展系である LS-VR-Q アルゴリズムは，C-table と呼ばれる頻度テーブルを各状態毎に定義して，その上で LS 価値関数を計算して行動を選択していた．そのため鮮明化や背景化が Q 値に対して行われるわけではなく，構造が複雑化された事で本来の満足化方策とは異なる性質を持つ．また，LS-Q 系アルゴリズムは行動選択方策に ϵ-greedy を用いなければならず，それのみでは探索が充分に行えずに学習を促進する事が出来なかった．ゆえに満足化の性質をそのまま保ち，乱数を用いずに学習を行う事が出来る RLLS は元々の LS の性質を引き継いだまま一般化した学習アルゴリズムであると言える．

6.7.1　設定

学習課題に用いるシミュレーションの物理環境は過去の応用案と同様にした [浦上 13]．腰のみを任意に動かす事が出来る大車輪ロボットが鉄棒に繋がれて静止した状態から学習は始まり，ロボットが可能な 3 つの行動，腰を “曲げる”，“延ばす”，“動かさない”，から逐次的に選択して，大車輪運動を獲得する事を目的とした課題である．エージェントが認識できる状態数は上半身の角度を 24，上半身と下半身のなす角度を 5，上半身の角速度を 7 に等分割した 840 種である．一回の行動選択と状態の変化を 1 step として，それを 110,000 step 行い，そのシミュレーションを 50 回行った結果を平均した．また状態は 1,000 step 毎に初期状態に強制的に戻される．報酬は初期状態であるロボットが垂直に下に向いている状態を角度 $\theta = 0$ として，step 毎に $r = \theta/\pi$ が与えられる．比較に用いるエージェントは高橋の研究で良い成績を有していた LS-VR-Q アルゴリズム (最も成績が良い $R = 0.8$ の場合 [高橋 13]) と，最も一般的な学習アルゴリズムとして Q 学習を用いる．上述のアルゴリズムの行動選択には ϵ-greedy を用い，ϵ は 1.0 から始まり，等間隔で徐々に減衰して 100,000 step の時点で 0.0 になるように設定する．本研究で提案する RLLS アルゴリズムは ϵ-greedy を必要としないため，最初から $\epsilon = 0.0$ に設定する．その代わり，全ての状態 s_i が持つ R_i 値は全て経験的な $R_i = 4.5$ に固定し，未来信頼度

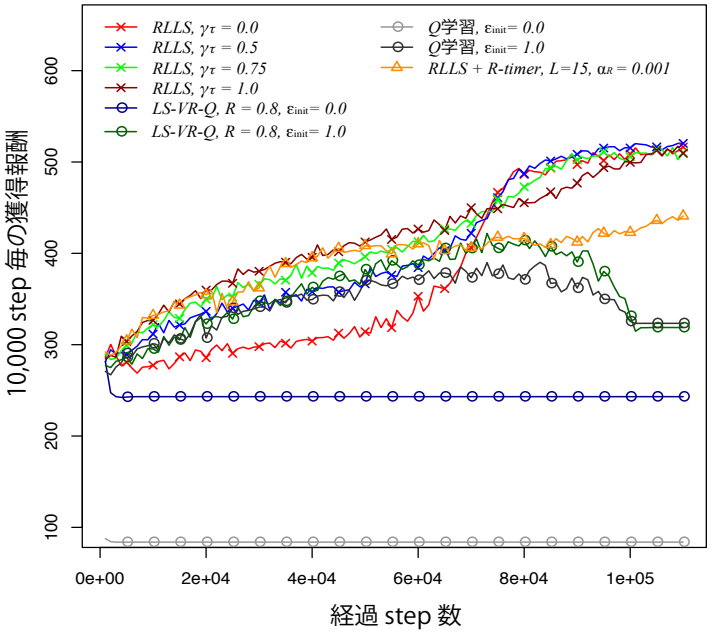

図 6.1: 獲得報酬の推移：学習率 $\alpha = 0.1$ の場合

割引率には $\gamma_\tau = \{0.0, 0.5, 0.75, 1.0\}$ を用いてそれぞれ比較した．また，動的な基準値の更新法についても言及するため，RLLS と R-timer を併用したアルゴリズム (R-timer RLLS) の検証も行う．L step 収益の長さは $L = 15$，馴化率 $\alpha_R = 0.001$ とした．また，乱数を用いない RLLS アルゴリズムとの比較のために， 最初から $\epsilon = 0.0$ に設定した LS-VR-Q アルゴリズムと Q 学習とも比較する．各アルゴリズムの割引率には $\gamma = 0.9$ を用いた．それに合わせ，R-timer RLLS の未来信頼度割引率には $\gamma_\tau = 0.9$ とした．

6.7.2 結果及び考察

シミュレーションの結果として，縦軸は初期状態に戻されるまでの 1,000 step 毎の報酬の総和の推移を図 6.1(学習率 $\alpha = 0.1$) と図 6.2(学習率 $\alpha = 0.9$) に示す．未来信頼度割引率 γ_τ はそれ以降に出現する行動系列を試行した強さを，どの程度その行動に対する信頼度変数に反映するかを意味する影響度であると解釈できる．シミュレーション結果には ϵ-greedy を用いて学習初期の探索を促さなければ学習が行えない事が示されている．それに対して RLLS は $\epsilon = 0$ というランダムな探索を全く行わない学習でも学習が行えていた．また学習率 $\alpha = 0.1$ の場合，未来信頼度割引率 γ_τ が低い時には，ある段階での獲得報酬の急激な上昇が見られるが， γ_τ が高くなる毎に，徐々に報酬が上昇していく傾向が見られる．学習率 $\alpha = 0.9$ の場合にも未来信頼度割引率 γ_τ に対する基本的傾向は変わっていないが，全体的な学習速度が圧倒的に早まっている．しかし最終的には LS-VR-Q 学習 に劣っている．これは決め打ちした RLLS の基準値 R がさほど良い基準でなかったためであると考えられる．しかしながら大車輪課題は多次元の物理量を離散化して状態認識しているので，そもそも学習率が高い方が行動を学習し易い．他方，RLLS は学習率 α や ϵ-greedy によるランダム探索等のパラメータに依存せず，決めうちの基準値 R でも一定以上の学習を行える点で，従来の C-table を導入して LS を実装したアルゴリズムより汎用性が高いと解釈できる．

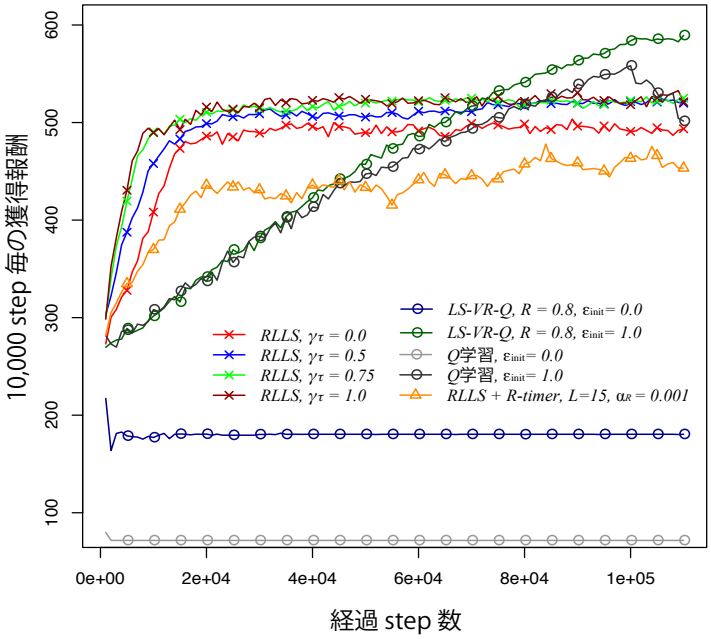

図 6.2: 獲得報酬の推移：学習率 $\alpha = 0.9$ の場合

R-timer を併用して動的に基準を求める RLLS アルゴリズム (R-timer RLLS) は予め持った基準値を与えられた RLLS には劣るものの，学習率 $\alpha = 0.1$ の場合では ϵ-greedy を用いずに学習できており，Q 学習や LS-VR-Q よりも良い成績を有している．学習率 $\alpha = 0.9$ のときは学習そのものはできているものの，予め持った基準値を与えられた RLLS にはやはり劣り，ϵ-greedy を用いた Q 学習や LS-VR-Q にも 40,000 step 前後で獲得報酬を抜かれてしまう．これは R-timer が良い基準値を獲得できなかったことが問題だと考えられる．しかし初期の報酬獲得の上昇の仕方等，予め持った基準値を与えられた RLLS と類似点を持っており，RLLS の ϵ-greedy を用いず，また自律的に探索を促すことができる性質が現れているといえる．

第7章　本研究の成果

本研究は定常・非定常かもわからない環境の中で意思決定するという前提下で，実際に複雑な環境下で選択している人間の意思決定傾向が有効に働く事を示した．具体的には LS を LSX, RLLS と段階的に拡張していき，広い問題範囲で扱える事を示した事が本研究の成果である．これは LS の拡張性の高さを示しており，現時点では離散的な LS の定義域を連続値へと拡張すれば，より広い意味でパラメータ空間を捉える事が可能になり，上記のより現実的な環境下における選択を柔軟にする示唆が得られた．

我々はまず人間の意思決定傾向，満足化を表現できる価値関数である LS を複数の選択肢，そして満足化に対して拡張した Extended LS (LSX) を考案した．オンライン *LSX* アルゴリズムでは無数の選択肢を持つバンディット問題を扱う事が可能で，かつ良い成績を導く事が出来る．しかし人間においてはワーキングメモリを始めとする他の認知的作用の影響の方が大きいため，その選択手法との直接的な比較を行う事はできない．そのため，現時点では *LSX* は人間の認知的な性質に習った数理モデルを用いた，非定常環境や膨大な選択肢を持つ課題に対して有用なアルゴリズムと扱っている．それらは認知的な背景を持つ特性ではあるものの，性質としては非常に単純であり，非論理的ではあるものの，より純化されたルールであると考えられる．即ち逆説的になるが，認知的性質に含まれる未知環境下での学習に有効な性質の幾つかを見いだし，かつその応用例を本研究において示すことが出来たと考えられる．特に本研究でその一端として示した，最適基準 *LSX* アルゴリズムは，最適な基準値を発見できたというだけの条件を与えた場合の *LSX* の性能を示すアルゴリズムであり，*LSX* の潜在的な性能が UCB1-tuned という既存で最も強い意思決定アルゴリズムの一つを上回る可能性を言及している．ここでの問題は基準値の設定にあり，本研究の単純なシミュレーションからも *LSX* の性能は基準値の学習に依存する事がわかる．統計的基準 *LSX* アルゴリズムによって，問題毎に適切な動的な基準値の更新法を与える事が可能である示唆が得られた．この課題に対しては LS の背景にある人間が選択を行う際，環境から得られた情報を如何にして活用し，基準値をどのように設定するかが解決の糸口になるかもしれない．

本研究では，定常・非定常に対する不確実さを持つ環境の下，柔軟に探索と利益追求をバランシングする意思決定における人間の特性を，より現実的な課題である強化学習に応用する事も目的としていた．しかし LSX は共変動情報を背景としているために実数スケールの価値関数への応用が出来ていなかった．我々は LSX に対する考察を基に，満足化傾向を直接的に残した実数スケールへの拡張モデルである RLLS を考案した．基準値を如何にして自然に獲得するかという，満足化の方策としての中枢的な問題を残したままではあるが，事前に，あるいは本研究で考案した R-timer 基準値更新等で，ある程度正しい基準値を付与する事が出来れば，強化学習全般でも満足化が有効である事を示した．また，強化学習における信頼度変数の伝搬に関する手法を考案し，それが学習の速さに関係する事も明らかになった．満足化における基準値はある種のエネルギーコストに対する目標値と見なす事が出来る．即ち，満足化が実装可能になる事により，より動物的なエージェントとして強化学習エージェントを自身の消費カロリーに釣り合う行動を見つけて，生存し続ける事を目的としたエージェントとして定義する事が可

能になったと言える．更に満足しているという均衡した状態から環境を学習し，目標値 (基準値) を動的に獲得できるアルゴリズムが開発されれば，より高度な知的活動エージェントとしての強化学習エージェントの発展を望む事が出来ると考えられる．

本研究は短期的には時間的に限られていたり非定常あるいは複雑な環境下での自律的意思決定への応用に寄与する．長期的には，いずれ人間が機械に依頼する課題はより現実的で複雑で非定常になっていく事を考慮し，より高度な知的エージェントの構築に際して，定常的な環境での最適化を前提としない本研究の成果が礎として貢献するものだと考えられる．

関連図書

[Manktelow 99] K. I. Manktelow., Reasoning And Thinking, *Psychology Press* (1999).

[Wickelgren 77] W.A. Wickelgren., Speed-accuracy tradeoff and information processing Dynamics, *Acta Psychologica* 41, pp. 67–85 (1977).

[Tenenbaum 11] J. B. Tenenbaum., C. Kemp., T. L. Griffiths., and N. D. Goodman., How to Grow a Mind: Statistics, Structure, and Abstraction, *Science*, vol. 331, no. 6022, pp. 1279–1285 (2011).

[Oaksford 94] M. Oaksford. and N. Chater., A rational analysis of the selection task as optimal data selection, *Psychological Review*, 101, pp. 608–631 (1994).

[Takahashi 10] T. Takahashi., M. Nakano. and S. Shinohara., Cognitive symmetry: Illogical but rational biases, *Symmetry: Culture and Science*, Vol. 21, No. 1-3, pp. 275–294 (2010).

[Hattori 07] M. Hattori. and M. Oaksford., Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis, *Cognitive Science*, 31, 5, pp. 765–814 (2007).

[篠原 07] 篠原修二, 田口亮, 桂田浩一, 新田恒雄, 因果性に基づく信念形成モデルと N 本腕バンディット問題への適用, **人工知能学会論文誌**, Vol.22, No.1, pp.58–68 (2007).

[Takahashi 10] T. Takahashi., M. Nakano. and S. Shinohara., Cognitive symmetry: Illogical but rational biases, *Symmetry: Culture and Science*, 21, 1–3, pp. 275–294 (2010).

[Takahashi 11] T. Takahashi., K. Oyo. and S. Shinohara., A Loosely Symmetric Model of Cognition, *Lecture Notes in Computer Science*, No. 5778, Springer, pp. 234–241 (2011).

[Kahneman 79] D. Kahneman. and A. Tversky., Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47(2), pp.263–292 (1979).

[Sutton 96] R. S. Sutton., Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding, *Advances in Neural Information Processing Systems 8*, pp.1038–1044, MIT Press (1996).

[Gigerenzer 02] G. Gigerenzer., Calculated risks: How to know when numbers deceive you, *Simon and Schuster*, New York (2002).

[Kohno 12] Y, Kohno. and T. Takahashi., Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, *SCIS-ISIS 2012*, Kobe Convention Center (Kobe Portopia Hotel), pp.1166–1171 (2012).

[Sutton 00] R. S. Sutton. and A. G. Barto., 強化学習, **森北出版**, (三上, 皆川 訳) (2000).

[大用 10] 大用庫智, 高橋達二, 因果帰納と意思決定を結ぶ緩い対称モデル, **第 *27* 回日本認知科学会**, P3–34 (2010).

[Tversky 74] A. Tversky. and D. Kahneman., Judgment under uncertainty: Heuristics and biases, *Science 185 (4157)*, 1124–1131 (1974).

[Kahneman 84] D. Kahneman. and A. Tversky., Choices, values and frames, *American Psychologist 39 (4)*, 341–350 (1984).

[Simon 56] H. A. Simon, Rational choice and the structure of the environment, *Psychological Review, 63*, 261–273 (1956).

[Uragami 11] D. Uragami., T. Takahashi., H. Alsubeheen., A. Sekiguchi., and Y. Matsuo., The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning, *Proceedings of the 2011 IEEE International Conference on Mechatronics and Automation*, August 7–10, Beijing, China , pp. 410-415 (2011).

[Auer 02] P. Auer., N. Cesa-Bianchi. and P. Fischer., Finite-time analysis of the multi-armed bandit problem, *Machine Learning*, 47, pp. 235–256 (2002).

[Wang 05] S. Gelly, Y. Wang., R. Munos. and O. Teytaud., Modification of UCT with Patterns in Monte-Carlo Go, *Technical Report*, No.6062, INRIA (2005).

[Kahneman 79] D. Kahneman. and A. Tversky., Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47(2), 263–292 (1979).

[Kohno 12] Y. Kohno. and T. Takahashi., Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012)*, Kobe Convention Center (Kobe Portopia Hotel), 1166–1171 (2012).

[甲野 14] 甲野佑, 高橋達二, 柔軟な意思決定機能のための認知特性の応用と検証, JSAI 2014(2014 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2N5-OS-03b-2 (2014).

[Kohno 14] Y. Kohno. and T. Takahashi., A Satisficing Strategy with Variable Reference in the Multi-armed Bandit Problems, *Proceedings of 12th International Conference of Numerical Analysis and Ap- plied Mathematics (ICNAAM 2014)*, (2014).

[Simon 56] H.A. Simon., Rational choice and the structure of the environment, *Psychological Review*, 63, 261–273 (1956).

[高橋 13] 高橋優太, 甲野佑, 高橋達二, 認知的な強化学習モデルに対する基準学習の応用と考察, JSAI 2013(2013 年度人工知能学会全国大会 (第 28 回)) 予稿集, 1L3-OS-24a-4in, (2013).

[浦上 13] 浦上大輔, 対称性推論と運動学習の分節化, LS モデルを応用した Q 学習による大車輪ロボットの実現, JSAI 2013(2013 年度人工知能学会全国大会 (第 27 回)) 予稿集, 1L3-OS-24a-5, (2013).

[太田 14] 太田 宏之, 甲野佑, 高橋達二, 線条体ニューロンの持続的発火と強化学習, 2014 年度人工知能学会全国大会, JSAI 2014 (2014 年度人工知能学会全国大会 (第 28 回)) 予稿集, 2N5-OS-03b-4, (2014) .

[大用 15] 大用庫智, 市野学, 高橋達二, 緩い対称性を持つ因果的価値関数の認知的妥当性と N 本腕バンディット問題におけるその有効性, 人工知能学会論文誌, 30, 2, 403–416, (2015).

[甲野 15] 甲野佑, 高橋達二, 満足化とその基準の動的な更新による強化学習の促進, JSAI 2015 (2015 年度人工知能学会全国大会 (第 29 回)) 予稿集, 2L1-1, (2015).

[Kohno 15] Y. Kohno. and T. Takahashi., A cognitive satisficing strategy for bandit problems, *International Journal of Parallel, Emergent and Distributed Systems*. (published online on Sep. 2, 2015), http://dx.doi.org/10.1080/17445760.2015.1075531, (2015).

[Hartland et al. 06] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, Multi-armed bandit, dynamic environments and meta-bandits, In Advances in Neural Information Processing Systems(NIPS-2006) Workshop, Online Trading Exploration Exploitation, 2006.